Genetic Analyses of Age at Onset Traits

Carl Anderson

Thesis submitted for the degree
of
Doctor of Philosophy

University of Edinburgh

2007

Declaration

I declare that this thesis was composed by myself and that the work contained herein is my own.

The work has not been submitted for any other degree or professional qualification.

Carl Anderson

Abstract

The identification of factors underlying complex trait variation is a major goal in the field of genetics. For normally distributed, fully observed trait data there are many well established statistical methods for partitioning phenotypic variation and for mapping quantitative trait loci (QTL). Survival or time-to-event traits often follow non-normal distributions and frequently contain partially-known (or censored) trait data. If standard statistical methods are used to analyse age at onset data a bias can be introduced through a failure to account for the non-normal distribution of the data and the presence of censoring. Complex statistical methods have been developed to partition trait variation and map QTL for age at onset or survival traits. In this thesis, the use of these survival analysis methods is compared to more established statistical methods for the analysis of age-at-onset data.

A brief introduction to the analysis of human variation and the issues associated with the analysis of age at onset data is given. The methods currently used to partition trait variation and map QTL for survival traits are discussed (Chapter 1). Age-specific penetrances can be used to model the age-at-onset of disease in unaffected individuals. This parametric method is used to identify loci underlying susceptibility to a novel co-morbid psychiatric phenotype (depression and unexplained swelling). The method is compared to a non-parametric variance component (VC) QTL mapping method that does not account for the age at onset of the disease. Parametric linkage analysis identified two suggestive

loci, neither of which were supported by the standard variance component analysis. VC analysis identified a suggestive linkage region on chromosome 14 which decreased upon fine mapping (Chapter 2).

Many of the current methods used to analyse survival data in human genetics are based on methods originally derived by animal geneticists. The analysis of survival traits in some experimental populations is simplified by the presence of fully inbred lines. However, for complex traits the methods are both computationally intensive and not widely available. A grouped linear regression method is proposed for the analysis of continuous survival data in fully inbred lines. Using simulation the method is compared to both the Cox and Weibull proportional hazards models and a standard linear regression method that ignores censoring. The grouped linear regression method is of equivalent power to both the Cox and Weibull proportional hazards methods, is significantly better than the standard linear regression method when censored observations are present and is computationally simple (Chapter 3).

A sample of 446 monozygotic (MZ) twins, 633 dizygotic (DZ) twins and 223 siblings was used to partition the inter-individual variance in age at menarche. The analysis was carried out using both a standard method which failed to account for the censored nature of the data and a mixed effects Cox model which fits a frailty model to the random effects. The standard methodology suggested that an additive genetic model best described the data. The most parsimonious model when using the frailty method included additive genetic and common environmental effects (ACE). The difference between the two models was caused by the different ascertainment of the siblings. The frailty model estimated the heritability of age at menarche to be 0.57 (Chapter 4).

In Chapter 5, a sample of 2,685 pseudo-independent sib-pairs is used in a genome-wide linkage scan for QTL underlying variation in age-at-menarche. The sample comprises of the adolescent sample discussed in chapter 4, and three adult cohorts. The proportion of censoring in the sample is 1.20% so a standard QTL mapping method is used. Two QTL of suggestive significance are identified on chromosomes 11p and 3p. The candidate genes WT1 and FSHB are located within the linkage peak on 11p. After the removal of bivariate outliers a locus on chromosome 12q was identified. No significant QTL were detected which suggests age-at-menarche is influenced by multiple genes of small effect. The thesis concludes with a general discussion (Chapter 6).

Acknowledgements

The work presented in the thesis was carried out over three years while registered as a postgraduate research student at the University of Edinburgh (UK). Throughout this time I was funded by a Medical Research Council (UK) postgraduate research studentship. I would like to thank the MRC for supplying me with the finances to undertake the reasearch presented herein.

The first 18 months of my research studies were spent at the Institute of Evolutionary Biology (formerly The Institute of Cell, Animal and Population Biology), University of Edinburgh, UK. The remainder of my PhD was spent at The Genetic Epidemiology Group, Queensland Institute of Medical Research, Australia. I am indebted to many people at these institutes and would like to take this opportunity to acknowledge those who have made the completion of this thesis possible.

I would like to thank my principal supervisor, Peter Visscher for his support and guidance throughout my PhD studies. I could not have completed this thesis without Peter's help, and I really appreciate the fact that his door was always open - even at 5.30pm on a Friday!

I would also like to thank my second supervisor, Douglas Blackwood, for introducing me to the field of psychiatric genetics and for supplying me with my first dataset. With regard to this project, I would like to thank Matthew

Dunnigan, Tony Pelosi, Val Murray, Irene McKee and George McDonald for their work ascertaining and phenotyping the participating families. In addition, I would like to extend thanks to Allan Maclean, David Burt and David Morrice for their genotyping work. For his help with the various censoring mechanisms described in chapter 3, I would also like to thank Ian White.

A great deal of thanks is owed to Nick Martin for permitting me to continue my PhD studies under Peter's supervision at QIMR. When I embarked on my PhD I could never have imagined that I would get the opportunity to work at a 'down-under' research group. It has been a once in a lifetime opportunity for me, and one which I have greatly enjoyed. I would like to thank Coral Pink, Carole Ferrier and Jayne Glendinning for making the change of hemispheres a smooth and pleasant one. In my time at QIMR I was able to work on one of the many available twin datasets, and for this opportunity I am very grateful to Nick, Ann Eldridge and Marlene Grace. I would also like to thank Gu Zhu, Scott Gordon, Dale Nyholt and David Duffy for practical help on many projects. I am eternally grateful to both Allan McRae and Stuart Macgregor, who together were always my first port of call with a programming or statistical problem. Special thanks also to Sue Treloar for her support throughout my 18 months in Australia.

To my fellow PhD students and friends, both at Edinburgh and Brisbane, I would like to say a huge thank you. I have many fond memories of my time as a PhD student and the majority of these involve either bizarre afternoon tea conversations, a drunken Friday night out (whether it be on The Cowgate or down The Valley) or a rare moment of footballing glory (and here I would specifically like to thank everyone I played with for Sporting ICAPB - surely the most mighty team to ever grace the King's Buildings).

I owe a great deal of gratitude to my family, and in particular my parents, for their continued support and belief throughout what has surely seemed a very long and protracted student career. Finally, I would like to thank Morven, without whose care and encouragement the completion of thesis would have been all the more difficult.

Publications

The following publication has resulted as a direct outcome of research described in this thesis:

A simple linear regression method for QTL linkage analysis with censored observations, **Anderson CA,** McRae AF, Visscher PM, Genetics **173**: 1735-1745 (2006)

The following papers are in final preparation, and have resulted as a direct outcome of research described in this theis:

A linkage study in families with major depression and co-morbid unexplained swelling (idiopathic oedema) **Anderson CA** Dunnigan MG, Pelosi AJ, Murray V, McKee I, McDonald G, Burt DW, Morrice DR, Muir WJ, Visscher PM, Blackwood DHR.

Estimation of variance components for age at menarche in twin families, **Anderson CA,** Duffy DL, Martin NM, Visscher PM

A genome-wide linkage scan for loci influencing variation in age-at-menarche, **Anderson CA,** Treloar SA, Montgomery GW, Martin NG, Visscher PM,

Table of Contents

Chapter 1

Introduction

1.1 Identifying factors underlying human variation

With over 6.5 billion people on the planet no two individuals are phenotypically identical. Even between identical twins most traits, from height to personality to disease susceptibility, show variation. The amount of variation between individuals differs from one trait to the next. What are the causes underlying this variation? The simplest answer is chance; any variation between two individuals is just random variation around the trait mean. This, however, does not explain the observation that, on average, two closely related individuals are more phenotypically similar than two unrelated individuals. Are there factors associated with relatedness that influence trait variation? Related individuals share both environmental and genetic factors to a greater extent than unrelated individuals. For a trait where variation is explained largely by genetic and/or common environmental factors, two related individuals will have, on average, a greater similarity than two unrelated individuals. Identical (monozygotic) twins, who share all genes in common, and non-identical (dizygotic) twins, who share on average 50% of their genes in common, can be used to partition the trait variation into genetic, common environment, and error (random) components (see Chapter 4).

For many traits, for example height, body mass index (BMI) or IQ, a large proportion of the variation between individuals is explained by genetic factors. Much work has been carried out to identify the genetic loci which underpin this variation. It is hoped that insights into the underlying biology of a trait will be gained by identifying genetic loci underlying trait variation. Identifying genetic loci which cause variation in human disease susceptibility has a large potential to improve our understanding of the disease, our treatment of the disease and our ability to identify those individuals at high risk. The ease with which one can map such genetic loci depends largely on the genetics underlying each trait. In genetical terms, traits can be broadly characterised into two groups, Mendelian traits and complex traits.

1.2 Mendelian traits

Mendelian traits are named after Austrian monk Gregor Johann Mendel (1822-1884) who is considered the 'Father of Genetics' for his groundbreaking research into the principles of heredity. They represent the simplest case of genetic causation and occur when phenotypic variation within a family is explained by the genotype at a single locus. The genotype at the locus is said to be both necessary and sufficient for the change in phenotype. Because Mendelian traits are fully influenced by the genotype at a single locus, they are dichotomous traits (i.e. either present or absent).

Mendelian traits are easily recognised because they show clear inheritance patterns in families. If the Mendelian trait is manifest to an equal extent in individuals with one or two copies of the trait allele, then that characteristic is said to be dominant. Typically, dominant traits have a distinctive pattern of transmission within families. An affected individual classically has at least one

affected parent (if the trait is sex linked then this is not necessarily the case). If the affected parent is heterozygous at the disease locus (i.e. the affected parent has both a wild type allele and a disease causing allele at the disease locus) and the other parent is unaffected, which is typically the case for rare conditions, then approximately half the offspring will display the trait. A classic example of an autosomal dominant trait is Huntington's disease [OMIM 143100].

If the Mendelian phenotype is present to a greater extent in homozygotes than heterozygotes then the trait is said to be semi-dominant or codominant. Waardenburg syndrome [OMIM 193500] is a semi-dominant trait which occurs in humans. Heterozygotes manifest the classic Waardenburg symptoms (Type 1) and homozygotes display a more severe form of the syndrome (Type 3) (Hoth *et al.*, 1993). If both affected parents are heterozygotes at the disease locus, then half of the offspring will be heterozygotes and have Waardenburg syndrome type 1, a quarter of the offspring will be homozygous for the disease allele and have Waardenburg syndrome type 3, and a quarter will be unaffected. A special case of codominant inheritance is additive or dose-dependent inheritance, which occurs when the phenotype displayed by the heterozygotes is halfway between that of the two homozygotes.

If the Mendelian characteristic is only expressed in individuals who carry two copies of the trait allele (i.e. are homozygous for the trait allele) then the phenotype is said to be recessive. Parents of affected individuals do not display the recessive trait because they are typically heterozygous at the trait locus. In the case of consanguineous marriages the chance of both parents carrying the same recessive disease allele is increased. Thus, the incidence of recessive traits is increased in cases of parental consanguinity. Cystic fibrosis [OMIM 219700] is a classic example of a recessive trait.

There are a number of complications that change the basic Mendelian pedigree patterns (Scriver and Waters, 1999; Weatherall, 2001; Badano and Katsanis, 2002). Occasionally, while the trait is still caused by the genotype at a single locus, the phenotype may display incomplete penetrance. Incomplete penetrance occurs when the probability of the phenotype given the genotype is less than 1 (i.e all carriers of the disease allele will not necessarily display the phenotype). The penetrance of a particular phenotype can change with age, because while the genotype is fixed at conception, the corresponding phenotype may not present itself until late into adulthood. Huntington's disease [OMIM 143100], a neurodegenerative disease, is a good example of a Mendelian disease with a variable age at onset.

1.2.1 Mapping Mendelian traits

The mapping of genetic loci requires polymorphic genetic markers which allow one to follow chromosomal segments through a pedigree. The identification of restriction-fragment length polymorphism (RLFP) markers (Botstein *et al.*, 1980) and highly polymorphic microsatellite (short tandem DNA repeat) loci (Tautz, 1989; Weber and May, 1989) during the 1980s led to increased interest in the identification of loci underlying Mendelian traits. Microsatellite markers are highly polymorphic, so a randomly selected individual is likely to be a heterozygote, and therefore informative for mapping genetic loci. Furthermore, both microsatellites and RFLPs are found frequently throughout the entire genome (Strachan and Read, 2003). Prior to the discovery of these markers, most disease mapping studies followed a functional cloning paradigm. Candidate genes were chosen based upon knowledge of their gene-product and their role in a disease pathway. Indeed, this was the method used to map phenylketonuria (PKU) [OMIM 261600] to the hepatic enzyme phenylalanine hydroxylase

(PAH) gene, one of the first Mendelian disease to be successfully mapped to a genetic locus (Woo *et al.*, 1983). However, the advent of highly-polymorphic genome-wide markers allowed a migration away from the hypothesis driven functional cloning paradigm and towards hypothesis-free positional cloning.

Positional cloning, initially termed 'reverse genetics', requires no knowledge of the biology underlying the trait (Collins, 1992; Collins, 1995). Loci are identified solely on the basis of their map position. In the early years, cytogenetics was used to identify individuals with translocations which resulted in Mendelian disease. Positional cloning would then be carried out along the relevant chromosomes to identify the causal locus. One of the first Mendelian disorders to be mapped in this way was Duchenne muscular dystrophy [OMIM 310200] (Monaco *et al.*, 1986). As the cost of genotyping marker loci decreased, it became possible for genome-wide scans to be carried out to detect linkage between a marker locus and a disease locus. As this technology became more widely available, the need for cytogenetic studies to identifying candidate chromosomes decreased. The first Mendelian disease to be fully mapped using only positional cloning was cystic fibrosis (Tsui *et al.*, 1985).

If one is attempting to map a Mendelian trait, the method of choice is parametric (model based) linkage analysis (Ott, 1992; Terwilliger and Ott, 1994). Linkage occurs when two genetic loci are not segregating independently. Linkage analysis attempts to detect co-segregation between a marker locus and a particular trait of interest. As the mode of inheritance (dominant, codominant, recessive) is clearly observed in the pedigrees, these models can be applied in the analysis to increase the power to detect linkage through the use of parametric analysis methods. Furthermore, if, through the analysis of the pedigrees, it is suspected that a locus with incomplete penetrance is underlying the variation in the trait,

this can also be modelled. The age-specific penetrance of the disease can also be included in the model to account for unaffected individuals who are not yet at risk (Chapter 2). Throughout this thesis, parametric linkage analysis is defined as a method for which a genetic model (for example, a particular set of disease allele frequencies q and penetrances f_0, f_1, f_2) is applied to the linkage analysis by the user, typically following the visual inspection of the pedigrees. Therefore, parametric methods are defined as those for which assumptions must made regarding the genetic effect of the susceptibility locus.

To date, almost 2000 Mendelian characteristics have been mapped to genetic loci, and almost 10,000 have been mapped to a particular chromosome (Online Mendelian Inheritance in Man [OMIM] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM, 25th October 2006). The ease with which a genetic locus can be mapped for a given trait, is directly related to the genotypic relative risk at that locus for the given phenotype. Therefore, due to their large phenotypic effect, Mendelian traits are the easiest to map. However, the majority of human phenotypic variation does not follow simple Mendelian inheritance patterns, and these phenotypes are referred to as complex traits.

1.3 Complex traits

The exact boundary between Mendelian and complex traits is vague; here a complex trait is defined as a trait which, at the population level, is not explained by the genotype at a single locus. These traits do not follow the simple Mendelian inheritance patterns discussed earlier. Some complex diseases, such as polycystic kidney disease [OMIM 173900] (Pei, 2003) or early-onset Alzheimer's disease [OMIM 104300] (Scott *et al.*, 2000), have locus heterogeneity (multiple

disease loci segregating in a population). Any one of these loci is sufficient for the occurrence of the disease phenotype, but for rare conditions it is likely that only one such polymorphism will be segregating within a single family. If this is the case, then a disease which is complex at the population level, can appear to be Mendelian at the family level. However, for common complex diseases with locus heterogeneity, there can be multiple disease alleles segregating within a family.

Only traits that, at the level of the individual at least, are explained by the genotype at a single locus have been discussed so far. However, a complex trait within one individual (and the population) can also be influenced by multiple genetic factors. If a complex trait is caused by a small number of moderate effect loci, that trait is said to be oligogenic. If a multitude of small effect genes underlie variation in the phenotype, that trait is referred to as polygenic. The presence of any one risk allele is insufficient for the presence of the phenotype in both oligogenic and polygenic traits. Retinitis pigmentosa [OMIM 268000] is an example of an oligogenic disorder that occurs in humans, the presence of the disease requires a mutant heterozygote at a locus in both the RDS and ROM1 genes (Kajiwara *et al.*, 1994). Such interaction between genetic loci is referred to as epistasis (the joint effects of the loci are greater than the sum of their independent actions). Not all oligogenic or polygenic traits display epistasis, the effects of different loci can also act additively.

Polygenic traits can be classified as either binary traits or quantitative traits. Binary traits indicate the presence or absence of a particular trait of interest (for example dwarfism or high blood pressure), whereas quantitative traits are measured on a continuous scale (for example height or diastolic blood pressure). Polygenic binary traits can also be modelled as quantitative traits if one assumes

that the phenotypic 'liability' is an unobserved continuous trait. It is assumed that the trait is present after some critical threshold of susceptibility has been surpassed (Falconer and Mackay, 1996).

Complex traits are often not only influenced by genetic effects. Environmental influences also underlie variation in most complex traits. Furthermore, interactions can occur between environment influences and alleles at particular loci. For example, a genetic predisposition may only be evident in the presence of certain environmental cues. For some genetic traits, individuals without any predisposing risk alleles can display the same phenotype, and these individuals are said to be phenocopies. Phenocopies are caused either by environmental effects or random chance.

The genetic architecture of most complex diseases is still largely unknown. For most traits only a few causal polymorphisms have been identified, and these are typically of large effect (e.g. the BRCA polymorphisms underlying breast cancer susceptibility in familial forms of breast cancer (Easton, 1999)). However, it would be wrong to conclude that complex traits are influenced by only a few genes of large effect. It has been suggested that currently identified susceptibility loci represent the 'low-hanging fruit' (i.e. the identified susceptibility variants have a more complete penetrance, larger effect size, and simpler allelic architecture then the majority of loci underlying complex traits) (Pritchard and Cox, 2002). It is likely that a multitude of small effect polymorphism influence variation in complex traits, but these are yet to be mapped.

1.3.1 Mapping complex traits

The mapping of complex traits is much more difficult than the mapping of Mendelian traits. Locus heterogeneity, incomplete penetrance, oligogenic or

polygenic loci, epistasis, gene by environment interactions and phenocopies can all hinder the search for linked genetic loci (Lander and Schork, 1994; Risch, 2000). In comparison to Mendelian traits, a much greater number of individuals is required to map loci underlying variation in complex phenotypes. The number of individuals required to map a locus is directly related to the phenotypic effect size of the locus, the smaller the effect size the greater the number of individuals required to map the locus.

Several strategies have been suggested to remove some of the complexities when mapping quantitative trait loci (QTL). Many researchers have suggested that choosing narrowly defined subsets of complex disease will result in the selection of more homogeneous individuals (i.e. increased genotypic homogeneity as a result of less phenotypic heterogeneity) (Rice *et al.*, 2001). It is hoped that, by selecting these narrowly defined subsets of complex disease, a group of families can be identified that are segregating a rare polymorphism of large effect (see Chapter 2). These polymorphisms will present with near-Mendelian inheritance patterns in carefully selected pedigrees. This method was put to good use in the mapping of two breast cancer genes, BRCA1 (Hall *et al.*, 1990; Miki *et al.*, 1994) and BRCA2 (Wooster *et al.*, 1994; Wooster *et al.*, 1995).

Others have suggested that rather than use end-phenotypes (disease status) in linkage analyses, efforts should be made to carry out linkage analysis on intermediate phenotypes (endophenotypes). For example, rather than testing for linkage to myocardial infarction, one may want to test for linkage to blood pressure or cholesterol level. The rationale here is that the endophenotypes are closer to the gene action than the disease end-point. Using brain oscillations in the beta frequency range (1328 Hz), Edenberg *et al.* (2004) identified the GABRA2 gene as a candidate gene for alcohol susceptibility. They conclude

that GABRA2 influences susceptibility to alcohol dependence by modulating the level of neuronal excitation.

In addition to the careful selection of individuals and phenotypes for mapping complex traits, special statistical methods must be used to detect linkage to complex traits. The parametric linkage method described earlier is no longer the method of choice because the mode of inheritance of the complex trait is usually unknown. Under the correct mode of inheritance, parametric methods have more power than non-parametric methods (Goldgar, 2001). However, non-parametric methods have relatively high power with all modes of inheritance (Clerget-Darpoux *et al.*, 1986; Kruglyak *et al.*, 1996). Most attempts to map complex traits are therefore carried out using non-parametric methods such as affected sib-pair methods, regression methods or variance components methods. Throughout this thesis, non-parametric methods are defined as methods for which no assumptions regarding the genetic effect of the risk alleles at the susceptibility locus are made. Rather than these parameters being fixed by the investigator they are explicit or maximized within the model.

Due to the complexities described above, and despite the large number of studies searching for genetic loci underlying variation in complex traits, few traits have been unequivocally mapped to genetic loci (Glazier *et al.*, 2002). Positional cloning linkage analysis was used to identify a susceptibility gene for Crohn's disease (Hugot *et al.*, 2001; Ogura *et al.*, 2001). Following linkage and association studies, a positional candidate gene approach was used to identify frameshift and missense mutations within the NOD2/CARD15 gene that were associated with Crohn's disease. Other complex traits for which susceptibility genes have been identified include Alzheimer's disease (ApoE) (Corder *et al.*, 1993), type 1 diabetes (INS) (Bell *et al.*, 1984), type 2 diabetes

(PPARγ, CAPN10) (Altshuler *et al.*, 2000; Horikawa *et al.*, 2000), asthma (ADAM33) (van Eerdewegh *et al.*, 2002) and macular degeneration (CFH) (Haines *et al.*, 2005; Klein *et al.*, 2005).

With many diseases, susceptibility is not constant throughout an individual's lifetime. For example, Huntington's disease susceptibility is directly related to the number of trinucleotide repeats at the Huntington's disease locus, the more repeats the earlier the onset of the disease. Further insights into the genetics underlying a trait can be gained by using the age at onset of a trait as the phenotypic data in a QTL mapping experiment. For example, it is perhaps more pertinent to use the age at onset of high blood pressure when searching for QTL than to use presence or absence of high blood pressure. The rationale here is that the genetics underlying early onset high blood pressure are different to those underlying late onset high blood pressure. Age at onset traits have a number of special features which differentiate them from standard traits.

1.4 Age at onset traits

Age at onset data are an example of survival data, which is the term used to describe data giving the length of time from a start point until the occurrence of a particular event (or end point). In genetic studies, the start time (t_0) typically represents the time at which an individual is recruited into a study. If the event of interest is death then the data are truly survival times. However, the event of interest can be any event which can be clearly defined and readily timed. For example, the event of interest could be pregnancy, pacemaker failure, the re-occurrence of pain following anaesthesia or the absence of disease symptoms following a given treatment. The survival study design typically involves recruiting a cohort of individuals and repeatedly following-up these individuals

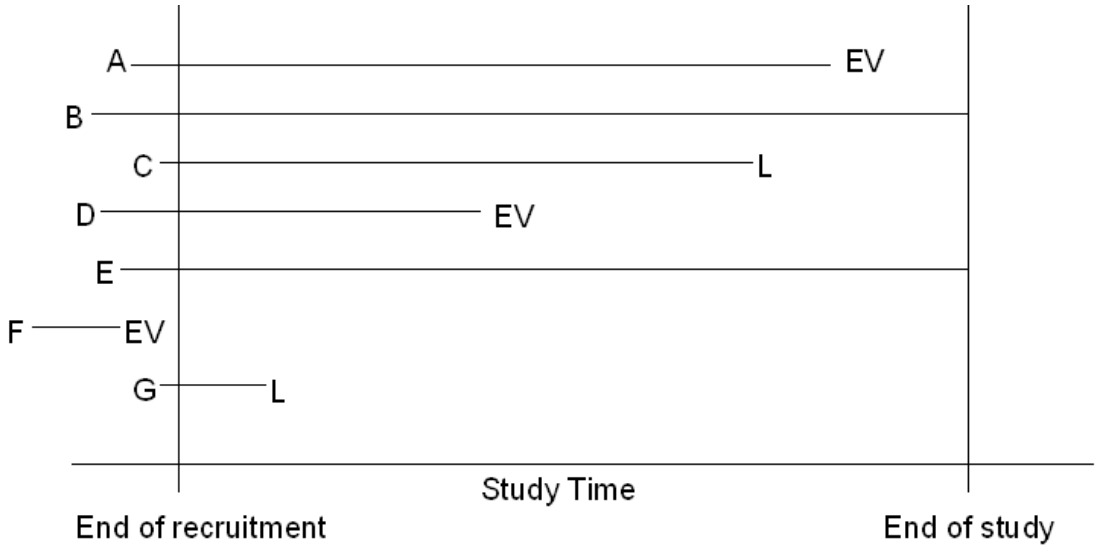

Figure 1.4.1: An example of 7 individuals (A to G) recruited to a survival study. EV = occurrence of the event of interest, L = lost to follow-up. Individuals A, D and F are uncensored and individuals B, C, E and G are censored. Individuals C and G are lost to follow-up and individuals B and E are censored because the event of interest had not occurred before the end of the study

until a previously defined point in time that marks the end of the study. At each interview the presence or absence of the event of interest is recorded. Once the event of interest has occurred for a given individual, that individual is no longer followed up. Figure 1.4.1 gives a typical example of 7 individuals from a survival study.

Survival data have several special features which separate them from standard quantitative trait data. Survival traits do not typically follow a normal or symmetrical distribution, and are frequently seen to be right skewed (i.e. positively skewed). In addition, the survival time of an individual may be only partially observed (i.e. the true time to event is unknown, it is only known that the event had not occurred at the time of last interview). The survival times of these individuals are said to be censored. In Figure 1.4.1, individuals B, C, E and G are censored. There are many reasons why an individual may be censored.

Firstly, individuals may be lost to follow-up, which indicates that they have left the study before the event of interest has occurred and before the end of the study period (individuals C and G in Figure 1.4.1 are lost to follow-up). For these individuals, it is only known that the event of interest had not occurred before they were lost to follow-up. Loss to follow-up can occur for many reasons including death (assuming the event of interest is not death) or the individual no longer wishing to participate in the study. Secondly, the study may end before a particular individual has experienced the event of interest, which is the case for individuals B and E in Figure 1.4.1. Due to the non-symmetrical distribution of the data and the presence of censored observations, special statistical tools are required to analyse survival data and map QTL for survival traits.

1.4.1 Mapping age at onset traits

If, due to a long period of follow-up and little loss to follow-up, the proportion of censoring in a sample is small then standard QTL mapping methodology can be used to analyse survival data. A transformation can be used to 'normalise' the distribution of survival times, which can remove the error introduced by using standard methodology to analyse survival traits. Alternatively, using the allele specific age at onset distribution of the uncensored individuals, the survival time of an individual censored at time x can be simulated. This method was implemented by Daw *et al.* (1999) to successfully map two previously known Alzheimer's disease susceptibility loci (PS1 and PS2) in a sample of 1,150 German individuals. This method is far from ideal as it assumes that all individuals will eventually experience the event of interest and this is not the case for many diseases, including Alzheimer's disease. When a reasonable proportion of individuals are censored, survival analysis methodology must be employed to analyse the data without introducing error. Several special 'survival analysis' methods have been developed to account for both the non-normal

distribution and the censored nature of the data when mapping QTL for age at onset traits. When compared to standard methods, survival analysis methods not only remove the bias introduced by the censored observations but also increase the power to detect linkage (Epstein *et al.*, 2003; Moreno *et al.*, 2005).

Many of the fundamental methods for mapping QTL for survival traits come from work carried out using experimental species and inbred lines. The mapping of QTL in these species is simpler than in outbred populations because the parental lines are usually fully inbred, and thus a maximum of two alleles are segregating at a given locus within the population. Several methods have been proposed for mapping QTL in inbred lines including those of Symons *et al.* (2002), Diao *et al.* (2004) and Diao and Lin (2005).

The mapping of QTL for survival traits in outbred populations is further complicated by the fact that survival times of individuals within a family are no longer independent. To model the correlation in survival times between individuals within a family, a correlated random effect (a frailty) must be used. Li (2002) proposed an additive gamma frailty model for linkage of age at onset traits. The gamma frailty model fits a gamma distributed random effect to the additive effect of the QTL in each founder in the population. The method is non-parametric and is therefore suitable for mapping QTL underlying complex age at onset traits. However, additional frailties must be calculated if one wishes to apply the method to extended families and this renders the maximum likelihood method intractable.

Subsequently, other methods have been suggested which follow the variance components framework typically used to map non-survival traits in outbred populations (Epstein *et al.*, 2003; Pankratz *et al.*, 2005; Diao and Lin, 2006).

As with standard complex trait analysis, parametric methods are not ideal because the distribution of the data is usually unknown. The methods of Diao and Lin (2005), Pankratz *et al.* (2005) and Diao and Lin (2006) are based upon the popular Cox semi-parametric model (Cox, 1972), and are therefore more suitable for analysing complex trait survival data. These methods, whilst more computationally intensive than the corresponding non-survival methods, can be used to map QTL and estimate variance components underlying age at onset or survival traits in outbred populations.

To date, most age at onset traits used in genetic linkage studies have been retrospective studies and have therefore not included censored data. Standard QTL mapping methodology is typically applied to these datasets. Perhaps due to the complexities involved in the analysis, few survival traits have been used in genetic mapping studies. Alcais *et al.* (2001) identified a locus on chromosome 6q underlying variation in wheezing age at onset in a sample of German Asthmatics.

1.5 Thesis Overview

This thesis aims to use age at onset information to identify loci underlying quantitative and disease traits. This is approached both directly through the use of age at onset data in QTL mapping experiments and indirectly by modelling the age-dependent penetrances in genome-wide scans for linkage. Methods which account correctly for age at onset data and age-dependent penetrances are compared and contrasted to standard methods. Where appropriate new methods are suggested for the analysis of survival data.

In Chapter 2, parametric linkage analysis is carried out on four families with major depression and unexplained swelling (a novel comorbidity). The age-

related disease susceptibility is modelled in the unaffected individuals to account for those individuals yet to enter the high risk age group. The parametric linkage analysis method (which accounts for the age related disease susceptibility) is compared and contrasted to a non-parametric linkage analysis method (which does not account for the age related disease susceptibility).

A novel method for linkage analysis of continuous age at onset data from inbred lines is proposed in Chapter 3. Because the method is based upon linear regression methodology, it can be readily implemented in many standard statistical packages. The new method is compared in terms of efficiency and power to those currently available.

In chapter 4, a biometric twin analysis is carried out on age at menarche data collected from a sample of Australian adolescent twins. This is an important first step when aiming to understand the factors underlying variation in a trait, and one which should be completed before attempting to map genetic loci. Variance components are estimated using both a survival analysis method and standard method. The results of the two methods are compared and discussed.

Age at menarche data was also available for three large Australian adult samples. There is no censoring in the adult sample and therefore the overall proportion of censoring in the combined adult and adolescent sample is small. In chapter 5, a standard variance components linkage analysis is used to identify loci underlying variation in age at menarche. A general discussion of the thesis and future direction of the field is provided in chapter 6.

Chapter 2

Genome-wide linkage analysis of four extended families with depression and comorbid unexplained swelling

2.1 Introduction

Recurrent major depressive disorder (MDD [OMIM 608516]) or unipolar depression, is a common psychiatric disorder predicted to rank second only to ischaemic heart disease as a source of disability world-wide by the year 2020 (Murray and Lopez, 1996). The US National Comorbidity Survey (NCS) reported a lifetime prevalence for major depressive episode of 21.3% in women and 12.7% in men (Kessler *et al.*, 1994). In a replication of the NCS, a measure of illness severity was included and the twelve month prevalence of MDD was estimated at 6.7%. Of these, 30% were rated as serious causing substantial disability in work and daily living or serious suicide intent (Kessler *et al.*, 2003; Kessler *et al.*, 2005).

Genetic factors have a substantial role in unipolar depression. Relatives of MDD sufferers are more likely than controls to develop illness and have a greater than twofold increase in relative risk (Sullivan *et al.*, 2000). Young offspring who have both a parent and a grandparent with depression are especially at

risk (Weissman *et al.*, 2005) and twin studies have consistently shown that genetic factors account for 40-70% of the risk for developing unipolar depression (Sullivan *et al.*, 2000). The risk of depression developing among family members is increased with early age at onset (under the age of 30) (Mendlewicz and Baron, 1981; Weissman *et al.*, 1984; Wickramaratne *et al.*, 2000; Kendler *et al.*, 2005) and a twin study of depressive symptoms in adolescence produced a heritability estimate as high as 0.79 (Thapar and McGuffin, 1994). Familial effects are also increased when depression is recurrent (Bland *et al.*, 1986; Kupfer *et al.*, 1989).

The mode of inheritance of MDD is unknown and a variety of genetic mechanisms are possible. No single genetic model is likely to explain the familial aggregation of all subtypes of the MDD phenotype. A quantitative meta-analysis of 5 twin studies concluded that familial aggregation could be explained by additive genetic effects (37%), and individual specific environmental effects (Sullivan *et al.*, 2000). Several segregation analyses have been carried out on families selected through MDD probands (Cox *et al.*, 1989; Marazita *et al.*, 1997; Maher *et al.*, 2002). The results of these segregation analyses have been inconsistent and inconclusive, perhaps because the results depend heavily on the definition of disease and population prevalence parameters, making them difficult to replicate between studies. A transmitted non-Mendelian recessive major effect locus with significant residual parental effect was the best model for a narrow definition of MDD. In the same study, a Mendelian co-dominant major effect locus with significant parental and spousal effect was the best fitting model under a more broad definition of the disorder (Marazita *et al.*, 1997). A recent study of pedigrees ascertained through a proband with recurrent early onset unipolar depression has reported genetic segregation consistent with a Mendelian dominant model. The presence within some families of a single major locus for

early onset depression seems likely (Maher *et al.*, 2002).

Genetic linkage analyses have been inconsistent, and many results have failed to be replicated across studies. The multifactorial aetiology of the disease and difficulties defining subsets of the disease are believed to be the main reason for the inconsistencies. Furthermore, in linkage studies, MDD has often been included with the bipolar affective disorders and used as a very broad definition of a mood or affective disorder (Venken *et al.*, 2005). Of the genome-wide linkage scans that selected probands based on an MDD phenotype, four have found regions of significant linkage. Abkevich et al (2003) included bipolar family members as affected and found linkage to 12q in males only. Holmans *et al.* (2004) found linkage to 15q and Camp *et al.* (2005) reported linkage on 3p and 18q. Nurnberger *et al.* (2001) included alcoholism in the phenotype definition and found linkage to 1p. Suggestive evidence of linkage to chromosome 1p was found following a genome scan by McGuffin *et al.* (2005) in a region which included the MTHFR gene previously associated with depressive symptoms.

The candidate gene CREB1, on chromosome 2q, was shown to be linked with early onset MDD (Zubenko *et al.*, 2002) and again by the same group in a follow-up genome scan (Zubenko *et al.*, 2003). No linkages to MDD have yet been replicated. Many other groups have looked for association or linkage of MDD to candidate loci but have failed to produce significant results (Balciuniene *et al.*, 1998; Serretti *et al.*, 2000; Zill *et al.*, 2002). It has been suggested that these inconsistent results can be reduced by selecting probands with a narrow definition of the disease, a similar severity, age at onset, gender and symptom profile (Rice *et al.*, 2001; Camp and Cannon-Albright, 2005). The aim is to reduce the impact of heterogeneity and environmental influences by selecting a homogeneous sub-sample of MDD.

Unexplained swelling symptoms (also known as idiopathic oedema or the fluid retention syndrome) are strongly associated with affective symptoms, including depression. Patients with idiopathic oedema exhibited significantly more evidence of affective disturbance, including major depression, than a group of female hospital outpatients (Pelosi *et al.*, 1986). The clinical phenotype comprises swelling symptoms involving the face, hands, breasts, abdomen and feet. The swelling worsens from morning to evening and is associated with a weight gain from 1-4 Kg. Objective evidence of pitting oedema is usually absent. Symptoms frequently present in post-pubertal females, and occasionally in pre-pubertal girls. Cases of swelling symptoms have occasionally been reported in men (Thorn, 1968; Edwards and Baylis, 1976; Streeten, 1978). Medical and physiological investigations exclude cardiac, hepatic, hypoproteinaemic and obstructive causes of oedema. Complement levels are normal, and features do not include urticaria, abdominal pain attacks or episodes of upper airway obstruction, thus distinguishing the condition from angioedema (Dunnigan and Pelosi, 1993). Unexplained swelling symptoms appear to be a component of a common syndrome characterised by affective symptoms (depression, irritability, anxiety, poor concentration, food cravings and sleep disturbance), somatic symptoms (fatigue, pain at several sites, headaches and visual blurring) and autonomic symptoms (flushing, urinary frequency, sweating, faintness, constipation and diarrhoea). Weight gain, obesity and a family history of swelling and diabetes mellitus are also associated with swelling symptoms. These affective, somatic and autonomic symptoms are shared by overlapping symptom clusters including fibromyalgia, chronic fatigue syndrome, irritable bowel syndrome, and pre menstrual syndrome. These conditions represent a common group of poorly understood syndromes that constitute a significant part of primary care and general hospital case load (Kroenke and Mangelsdorff, 1989).

A recent study examined the prevalence of unexplained swelling symptoms in the community (Dunnigan *et al.*, 2004). The study comprised 198 women attending a fracture clinic (median age 45, range 17 to 82 years), and 201 women attending their general practitioner (median age 40, range 16 to 77 years). The self reported prevalence of mild to severe swelling symptoms in the previous month was 33% and 28%, respectively. A logistic regression model of risk factors associated with swelling symptoms was applied to a dataset derived from a fracture clinic sample (n=198) and women attending a menopause clinic (n=201). Severe affective symptoms (Relative risk (RR) 43, 95% Confidence Interval (CI) 16-112, P <0.001) and moderate affective symptoms (RR 7.8, 95% CI 4-15, P <0.001), a family history of swelling symptoms (RR 4.5, 95% CI 2.3-8.8, P <0.001) and a body mass index (BMI) > 25kg/m^2 (RR 4.8, 95% CI 2.5-8.9, P <0.001) were all significantly associated with mild to severe swelling symptoms assessed by visual analogue scales. In summary, affective symptom severity provides the principal independent contribution to swelling symptom risk. BMI > 25 and a family history of swelling symptoms provide smaller independent contributions. The nature of the mechanisms underlying these associations remains uncertain. The aim of this present study is to identify loci underlying susceptibility to MDD using a narrowly defined subset of the disease. A genome-wide scan for linkage is performed on four unrelated families whose affected members demonstrate the typical phenotype for unexplained swelling comorbid with major depression disorder.

2.2 Methods

2.2.1 Family ascertainment

Four families with multiple cases of depression and unexplained swelling are described in Figure 2.2.1. The study was approved by the appropriate Multicentre

Research Ethics Committee and all participants gave full written informed consent. Families were ascertained by Dunnigan *et al.* (2004) in the course of a large study of the prevalence of unexplained swelling symptoms in hospital and general practice clinics. A branch of family 224 was first described by Dunnigan and Pelosi (1993). An operational clinical diagnosis of unexplained swelling was based on a description by the patient of intermittent swelling or bloating at two or more sites (face, abdomen, breasts, hands and feet) over at least six months, and usually for several years. Clinical examination with appropriate investigations, excluded hereditary angioedema, fibromyalgia and cardiac, renal, obstructive and hypoproteinemic causes of swelling. All family members were examined in the clinic or at home and a diagnosis of unexplained swelling made using the same operational criteria. Home interviews were also carried by a trained psychiatrist and a psychiatric research nurse using The Schedule for Affective Disorders and Schizophrenia. Diagnoses according to DSM-IV criteria (American Psychiatric Association, 1994) were reached by consensus between the interviewing psychiatrist and two other psychiatrists. From a total of 47 affected individuals, 28 had both MDD and unexplained swelling, while 8 had MDD only and 11 had unexplained swelling only.

2.2.2 Genotyping and error checking

Venous blood samples were obtained and genomic DNA was extracted from the peripheral lymphocytes using a standard salting out procedure by Genovar Biosciences (Miller *et al.*, 1988). PCRs were performed in a 12μl reaction volume containing 2.5pmol of each primer, 1 x Sigma PCR buffer, 0.2mM dNTPs, 0.25 units of Sigma Taq and 20ng of genomic DNA. The ABI Prism Linkage Mapping Set-MD10 (version 2) primers were used (average heterozygosity 79%). One of three dyes (FAM, HEX and NED) were used to label 312 microsatellites, and organised into 28 panels for running on automated sequencing machines.

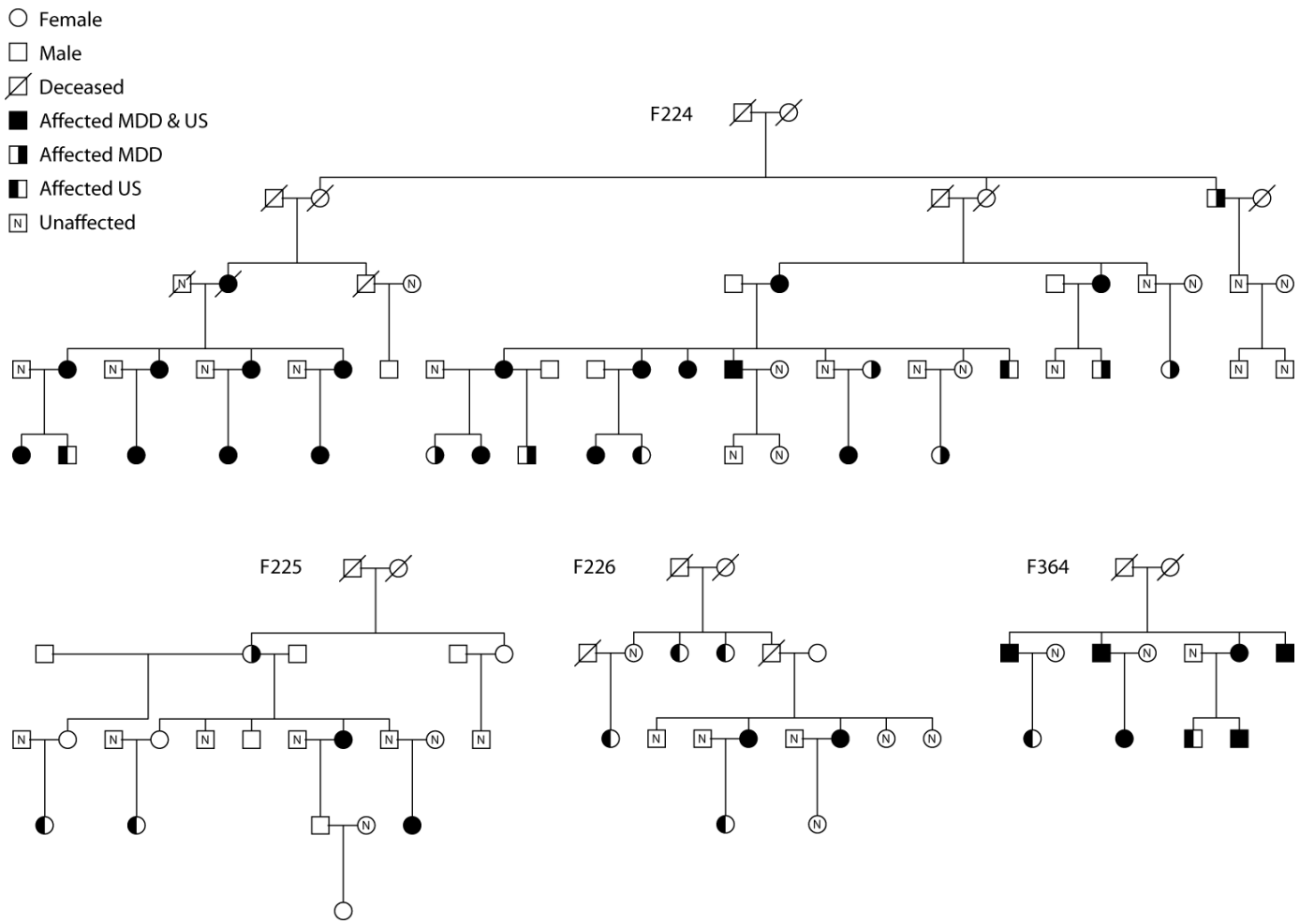

Figure 2.2.1: Pedigree structure of four Scottish families ascertained with depression and comorbid unexplained swelling.

The average marker density was 12cM. An MJ Research PTC-225 DNA Engine Tetrad (MJ Research) was used to amplify DNA fragments with 35 cycles of 20 seconds at 94°C, 30 seconds at 55°C, and 1 minute at 72°C. This was preceded by an initial denaturation step of two minutes at 94°C and ended with an extension step of 30 minutes at 72°C. PCR products were pooled for each panel and run in an ABI 377 automated DNA sequencer or ABI 3730 DNA Analyser with either 400HD rox or GS500 liz size standard. Allele sizes were determined using either GENOTYPER (3.6) or GENEMAPPER (3.0) analysis software (Applied Biosystems), then subsequently checked for quality manually. All genotyping was carried out by Alan Maclean at the Molecular Medicine Centre, University of Edinburgh, UK.

The genotypes were checked for inheritance errors using PEDCHECK (O'Connell and Weeks, 1998). Marker genotypes producing Mendelian segregation errors were removed from the study by re-classifying the marker genotypes of the relevant family members as unknown. The ABI marker map was used because this was calculated using a greater marker coverage than the present scan, and therefore provides better estimation of intermarker distance. A *flips* analysis was implemented through CRIMAP (Lander and Green, 1987) to check the given marker order was the most likely. A difference of -3 between the $\log_{10}$likelihoods of the original and hypothesized marker order was deemed significant enough evidence to reject the original marker order and remove the marker from further analyses. This level of significance equates to the hypothesized marker order being at least 1000 times more likely than the original marker order.

2.2.3 Parametric linkage analysis

Single-point parametric linkage analysis was carried out across all autosomes and the X chromosome using LINKAGE (Lathrop *et al.*, 1984). Parametric

linkage analysis has proven to be a powerful technique for mapping Mendelian diseases with well defined modes of inheritance (Tsui *et al.*, 1985). As a mode of inheritance must be specified for parametric linkage analysis its application in the analysis of complex traits is not straightforward. Greenberg *et al.* (1998) have shown that with a limited number of simple models, both dominant and recessive, parametric analysis can be successfully applied to complex disease. For single-point analysis, model misspecification can result in an upward bias in the estimation of the recombination fraction, while the strength of the reported LOD score remains unbiased provided dominance is correctly specified (Clerget-Darpoux *et al.*, 1986; Risch and Giuffra, 1992).

Two definitions of disease were used throughout the study to better model the comorbidity of the phenotype. The narrow definition defined individuals as affected if they were diagnosed with MDD, regardless of unexplained swelling affection status. The broad definition extended the narrow definition to include those individuals with only unexplained swelling (i.e. individuals with MDD and/or unexplained swelling are given as affected). An autosomal dominant and an autosomal recessive model were fitted under each disease definition, totalling four models for parametric linkage analysis. For the dominant and recessive models, disease allele frequencies of 0.012 and 0.300 were assumed, respectively. These are similar to those suggested from simulation (Pal *et al.*, 2001). Allele frequencies for all markers were set equal, i.e., the frequency for any given allele was set to $1/n$, where n is the total number of observed alleles. For each model, unaffected individuals were classified into one of five groups ($<$20yrs, $<$30yrs, $<$40yrs, $\geq$40yrs, and 'married in') to allow the age-dependent penetrance of MDD and US to be modelled (see chapter 1). The penetrance parameters used in the parametric linkage analysis are shown in Table 2.2.1. The model parameters were provided by Professor Douglas Blackwood of the University of Edinburgh

and are similar to those parameters used in the analysis of other psychiatric traits.

A series of marker-specific LOD scores were produced across all 23 chromosomes. A LOD score greater than or equal to 3.3 was accepted as genome-wide significance, which corresponds to a genome-wide false-positive rate of 0.05 or an asymptotic *P*-value of 0.000049, thus accounting for the multiple testing taking place in genome-wide scans for linkage (Lander and Kruglyak, 1995).

2.2.4 Further genotyping

A further 59 microsatellite markers from the ABI Prism Linkage Mapping set MD10 (version 2) were genotyped. The additional genotyping was carried out in regions of low marker information content. In total, 371 markers were spaced at an average density of approximately 10cM. The data were again checked for genotyping errors using PEDCHECK and CRIMAP (O'Connell and Weeks, 1998; Lander and Green, 1987). All identified errors were removed from the analysis and single-point parametric linkage analysis was repeated using the same models as in the initial analysis. Marker allele frequencies were estimated from all family members using Merlin (Abecasis et al 2002).

2.2.5 Affecteds only parametric linkage analysis

The modelling of complex diseases in parametric linkage analysis is difficult because penetrance, disease allele frequencies, population prevalence and phenocopy rates are uncertain. It is especially difficult to provide age-specific disease penetrances for unaffected individuals when the disease onset and susceptibility has not been well characterised. An affecteds only analysis, which models the disease only in affected individuals, was implemented using the same disease parameters as in the full analysis. Unaffected individuals had their disease status changed from unaffected to unknown. By comparing the results of this

Table 2.2.1: Parametric linkage penetrance parameters

Model	Genotype	Unaffected Penetrances					Affected Penetrances		
		(0 - 19yrs)	(20 - 29yrs)	(30 - 39yrs)	(≥40yrs)	Married In	MDD	US Only	Unknown
A	XX	0.0160	0.0240	0.0380	0.0465	0.0465	0.0005	0.0465	0.5000
A	Xx	0.2500	0.3900	0.6100	0.7500	0.7500	0.2900	0.7500	0.5000
A	xx	0.2500	0.3900	0.6100	0.7500	0.7500	0.2900	0.7500	0.5000
B	XX	0.0030	0.0050	0.0080	0.0100	0.0100	0.0001	0.0100	0.0465
B	Xx	0.0540	0.0840	0.1300	0.1600	0.1600	0.0620	0.1600	0.7500
B	xx	0.0540	0.0840	0.1300	0.1600	0.1600	0.0620	0.1600	0.7500
C	XX	0.0003	0.0012	0.0013	0.0013	0.0013	0.0013	0.0013	0.5000
C	Xx	0.0003	0.0012	0.0013	0.0013	0.0013	0.0013	0.0013	0.5000
C	xx	0.1500	0.6200	0.7000	0.7000	0.7000	0.7000	0.7000	0.5000
D	XX	0.0003	0.0012	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
D	Xx	0.0003	0.0012	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
D	xx	0.1500	0.6200	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000

MDD = Major depressive disorder, US = Unexplained Swelling. A = narrow-dominant model, B = broad-dominant model, C = narrow-recessive model, D = broad-recessive model. X = Wild-Type allele, x = Disease susceptibility allele. For the dominant and recessive models, disease gene frequencies of 0.03 and 0.12 were used, respectively.

analysis to the 'full' analysis, the effect of including the unaffected individuals in the model and modelling the age-specific disease penetrances in these individuals can be seen. An alternative approach would be to use a single set of penetrance parameters for the unaffected individuals (i.e. do not model the age-dependent penetrances in the unaffected individuals). This method has the advantage that the unaffected individuals remain in the analysis. The disadvantage of this method is that one still needs to derive a set of penetrance parameters for unaffected individuals, and these are difficult to obtain.

2.2.6 Non-parametric variance components analyses

A genome-wide variance component linkage analysis was carried out using the two step approach of George *et al.* (2000), which has been shown to detect QTL for complex psychiatric disorders (Visscher *et al.*, 1999). This method does not specify a genetic model for the trait, but identifies regions that explain a significant proportion of the phenotypic variation. The two definitions of disease used during the parametric analysis were again applied. Phenotypes were coded as a binary trait, with 1 for affected and 0 for unaffected. Identity-by-descent (IBD) coefficients between all genotyped individuals were calculated by the Markov Chain Monte Carlo method using LOKI (Heath, 1997). A simple linear model was assumed with regard to the MDD phenotype. Restricted maximum likelihood analysis (Lynch and Walsh, 1998) was performed using ASREML (Gilmour *et al.*, 2002). REML estimates variance components whilst taking account of the number of degrees of freedom used by fitting fixed effects, and is equivalent to ANOVA for balanced designs. Restricted maximum likelihood (REML) and maximum likelihood (ML) only differ when many fixed effects are included in the model. Given that no fixed effects are included in the present model, and that asymptotically the same distributions of the test statistic apply for both methods, there would be essentially no difference between REML and

ML for the present analysis. For ease of use the REML method was chosen because it is packaged with LOKI, through a portal provided by QTL EXPRESS (Seaton *et al.*, 2002).

A secondary analysis was carried out on the region showing the most significant linkage in the genome-wide scan. A threshold model was fitted using SOLAR (Duggirala *et al.*, 1997; Almasy and Blangero, 1998). The model takes into account the dichotomous nature of the phenotype. SOLAR estimates IBD probabilities using an algorithm based on applying regression formulae to results from a marker specific IBD estimation algorithm. The SOLAR estimation method has been shown to be less efficient (Sobel *et al.*, 2001), and therefore SOLAR was not used for the genome-wide variance component linkage scan.

2.2.7 Further genotyping in peak linkage region

Additional markers were added across the region with the largest LOD score to maximise multipoint marker information content. In total, 16 microsatellite markers and 20 SNPs were added across the linkage region. Several of the additional microsatellite markers and the entire additional SNP marker set were not placed on the ABI marker map. To obtain an accurate integrated map on which all markers could be placed the Oxstats map was used (Kong *et al.*, 2002; Myers *et al.*, 2005; Duffy, 2006). Markers were checked for genotyping errors using PEDCHECK and CRIMAP. Erroneous markers were removed from the analysis. Multipoint variance component analysis was carried out assuming a threshold normal model in SOLAR.

2.3 Results

2.3.1 Genotyping and error checking

Following CRIMAP *flips* analysis (Lander and Green, 1987) an alternative position for marker D2S364 was given that was approximately 17,000 times more likely than the one suggested by the ABI map. This marker was removed from further analysis. CRIMAP *flips* analysis was re-run across chromosome two, and no further markers gave a likelihood lower than -3.00. The given marker order was accepted as a true representation of the marker sequence.

2.3.2 Parametric linkage analysis

The preliminary single-point linkage analysis identified 93 markers with a LOD score greater than or equal to one (results not shown). An additional 59 markers were placed in regions of low marker density on chromosomes 2, 3, 6, 7, 8, 10, 11, 12, 17, 19 and 20. Following the additon of further markers and the use of estimated allele frequencies, 23 markers gave a LOD score greater than or equal to 1.00. Only two of the additional markers gave a LOD $\geq$ 1.00. Table (2.3.1) shows all markers with LOD score $\geq$ 1.00 following the addition of extra markers. Table 2.3.3 shows 'by family' LOD scores for the 5 markers with the highest single-point LOD scores identified by parametric linkage analysis. The markers where LOD $\geq$ 1.00 were spread across 14 different chromosomes, with the maximum single-point LOD score of 2.02 (marker D8S260) located on chromosome 8q. This is above the genome-wide level for 'suggestive linkage' (LOD = 1.9, $P = 0.0017$) put forward by Lander and Kruglyak (1995).

Chromosome 14 contains a region of further interest because three markers (D14S275, D14S288, D14S63), all within a 47cM interval spanning 14q11.2-

Table 2.3.1: Parametric marker-specific linkage results for markers showing a LOD score ≥ 1.00 in at least 1 family

Marker	Model	LOD score	Family
D1S2841*	Broad-dominant	1.17	364
D18S2800	Broad-recessive	1.01	226
D2S125	Narrow-recessive	1.37	ALL
D2S125	Broad-recessive	1.33	ALL
D4S2952†	Broad-dominant	1.10	226
D4S1595†	Broad-dominant	1.10	226
D4S1535	Broad-dominant	1.41	226
D4S2952†	Narrow-recessive	1.25	ALL
D4S1595†	Broad-recessive	1.13	226
D5S471	Narrow-recessive	1.11	ALL
D5S471	Broad-recessive	1.19	226
D7S516	Broad-recessive	1.20	ALL
D8S264	Narrow-recessive	1.15	ALL
D8S260	Narrow-recessive	*2.02*	ALL
D8S260	Broad-recessive	1.34	ALL
D10S1653*	Broad-dominant	1.19	225
D11S904	Narrow-dominant	1.01	224
D11S904	Broad-dominant	1.01	224
D11S968	Narrow-recessive	1.03	224
D11S968	Broad-recessive	1.21	224
D12S324	Broad-dominant	1.04	226
D12S324	Broad-recessive	1.17	226
D14S275	Narrow-dominant	1.73	ALL
D14S275	Broad-dominant	1.55	ALL
D14S288	Broad-recessive	1.05	364
D14S63	Broad-recessive	1.32	226

Table 2.3.2: Continued...

Marker	Model	LOD score	Family
D16S3046	Broad-recessive	1.38	226
D16S3068	Broad-recessive	1.06	226
D17S1857	Broad-dominant	1.25	ALL
D17S1868	Broad-dominant	1.21	ALL
D20S107	Narrow-dominant	1.01	ALL
D20S107	Narrow-recessive	1.68	ALL
DXS1214	Narrow-dominant	1.70	ALL

ALL indicates a LOD score summated across all families. Markers giving a LOD score $\geq$ 1.00 when using estimated allele frequencies but not with equal allele frequencies are marked by *. Markers added in the second genotyping stage are indicated by †. The maximum LOD score is shown in italics.

14q23.3, produced a LOD score greater than or equal to a 1-LOD threshold. Marker D14S70, which lies between D14S275 and D14S288, is the only marker in this region not to show a LOD $\geq$ 1. There is no evidence against linkage at this marker (LOD = -2), and the maximum LOD score at this locus is 0.56 under the broad-recessive model. The maximum LOD score for marker D14S275 was seen under the broad-dominant model, whereas for markers D14S288 and D14S263 the broad-recessive model produced the maximum LOD score.

2.3.3 Affecteds only parametric linkage analysis

'Affecteds only' parametric linkage analysis identified two further markers (D2S325 and D7S510) with a LOD score greater than 1.00. Of the 30 markers with a LOD $\geq$ 1.00 in the 'full' analysis, 10 produced a LOD score greater than or equal to 1.00 in the affecteds only analysis. A summary of the 'affected only' LOD scores is given in Table 2.3.4. Given that the linkage information provided by the unaffected individuals has been removed, it is not surprising to see a reduction

Table 2.3.3: Maximum LOD scores per-family for each of the five markers with the highest LOD scores from marker-specific parametric linkage analysis

	D8S260	D14S275	DXS1214	D20S107	D14S275
Family	Model C	Model A	Model A	Model C	Model B
F224	0.85	0.91	*1.22*	0.45	0.92
F225	0.11	0.17	0.03	-0.02	0.30
F226	0.58	0.46	0.45	0.57	0.03
F364	0.47	0.19	0.00	0.68	0.29
Total	*2.02*	*1.73*	*1.70*	*1.68*	*1.55*

Model A = Narrow-Dominant, Model B = Broad-Dominant, Model C = Narrow-Recessive. LODs greater than 1.00 are italicised.

in the number of LOD scores above a LOD-1.00 threshold. It is encouraging to note that of the top 8 markers with LOD scores ≥ 1.00 under the 'full' analysis, 7 have LOD scores equal to or greater than 1.00 under the 'affecteds only' analysis. This suggests that modelling the age-specific disease penetrance in unaffected individuals introduced little bias to the results.

2.3.4 Non-parametric linkage analysis

Following the single-point parametric linkage analyses, a genome-wide non-parametric variance component multipoint analysis was carried out. Results are shown in Figure 2.3.1. Heritability estimates for MDD only and MDD or unexplained swelling were 56% and 43%, respectively. Suggestive linkage peaks are seen on chromosomes 7 and 14, under the narrow and broad definitions, respectively. The linkage peak of 2.10 observed on chromosome 7 from variance component analysis is approximately 28cM away from marker D7S516, which gave a LOD score of 1.2 in the parametric linkage analysis (broad-recessive model). Following the preliminary parametric genome scan further markers were added

Table 2.3.4: MLINK marker-specific linkage results for markers showing a LOD score ≥ 1 in at least 1 family following 'affecteds only' analysis

Marker	Model	LOD score	Family
D1S2841	Broad-dominant	1.14	364
D2S125	Narrow-recessive	1.12	ALL
D2S325*	Broad-recessive	1.43	ALL
D4S1535	Broad-dominant	1.00	226
D4S2952	Narrow-recessive	1.04	ALL
D7S510*	Broad-dominant	1.10	ALL
D8S260	Narrow-recessive	1.35	ALL
D13S265	Narrow-recessive	1.23	ALL
D14S275	Broad-dominant	1.00	ALL
D14S288	Broad-recessive	1.05	364
D20S107	Narrow-dominant	1.00	ALL
D20S107	Narrow-recessive	*1.67*	ALL
DXS1214	Narrow-recessive	1.13	ALL

ALL indicates a LOD score summated across all families. Markers giving a LOD score ≥ 1 under the 'affected only' analysis but not in the 'full' analysis are marked by *. The maximum LOD score is shown in italics.

around D7S516. Markers D7S2427 and D7S519 are located under the variance component linkage peak on chromosome 7. Neither marker D7S2427 nor D7S519 gave a LOD score greater than 1 under the parametric linkage analysis, with the maximum observed single-point LOD scores being 0.97 (narrow-dominant model) and 0.66 (broad-dominant model), respectively. The most interesting region highlighted from the parametric linkage analysis was a 47cM region on 14q, which contained three markers with a LOD score ≥ 1.00. This finding is supported by variance component linkage analysis, which gave a broad peak across chromosome 14q, with a maximum LOD score of 2.68 (broad model) approximately 34cM from the centromere. Under the narrow definition of disease the maximum LOD score

was 2.50. Suggestive evidence for linkage was also found after fitting a threshold model with Solar (LOD = 2.17). Sixteen microsatellite markers and 20 SNPs were added across the linkage region on chromosome 14 to maximise the multipoint marker information content in the region . A multipoint variance components analysis was again carried out using SOLAR and a threshold model applied. The maximum multivariate-normal multipoint LOD score after the inclusion of the new markers was 1.56. Figure 2.3.2 shows the evolution of the chromosome 14 LOD score profile through the three stages of the study. The evidence for linkage reduced substantially when the information on identity-by-descent was maximised.

2.4 Discussion

This is the first genome-wide linkage study of major depressive disorder to use a narrow definition of MDD classified by a comorbid non-psychiatric trait (unexplained swelling). Suggestive evidence of linkage was found on chromosome 8q, which has previously been shown to contain bipolar affective disorder predisposition locus (Cichon *et al.*, 2001). Whilst the genetic links between bipolar and major depression remain unclear it seems unlikely that the locus identified in the present study is the previously identified bipolar locus. The deCODE map places the bipolar locus at 124.62cM and the marker identified in this study at 73.6cM. Both the 8q and 7p linkage regions identified in this current genome scan are novel loci for major depressive disorder. On chromosome 14 an initial variance component linkage peak of 2.60 (broad model) was found at a position of 34cM from the p-terminal. Linkage at the region is well supported by several analytical methods, with LOD $\geq$ 1 under parametric analysis using narrow-dominant, broad-dominant, and broad-recessive models. Positive linkage was found at this region in all four families. Following the

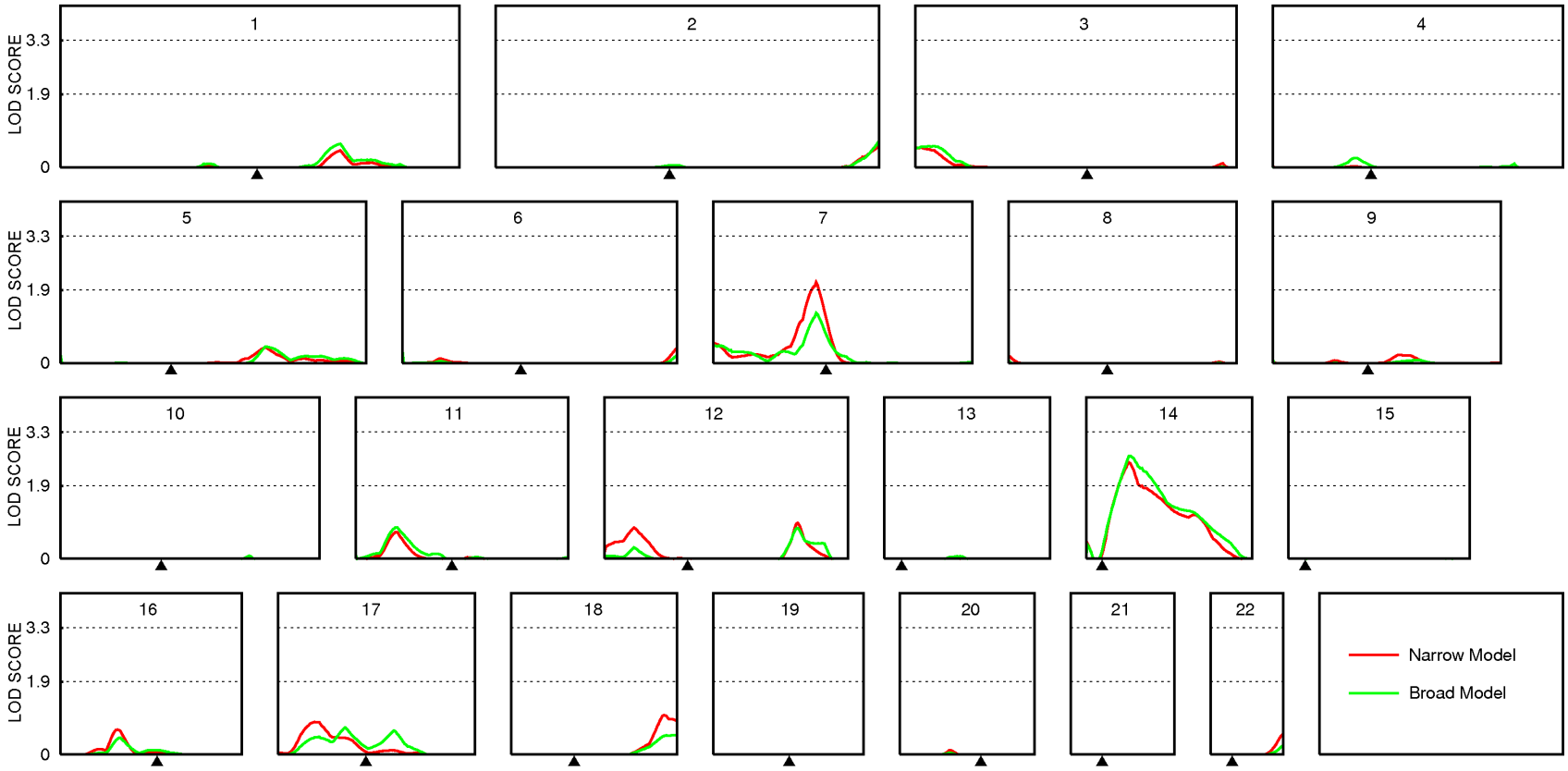

Figure 2.3.1: Variance component LOD scores by chromosome, calculated using MERLIN. Narrow model = Major depressive disorder regardless of unexplained swelling status. Broad model = Major depressive disorder and/or unexplained swelling. The centromere position is given by ▲.

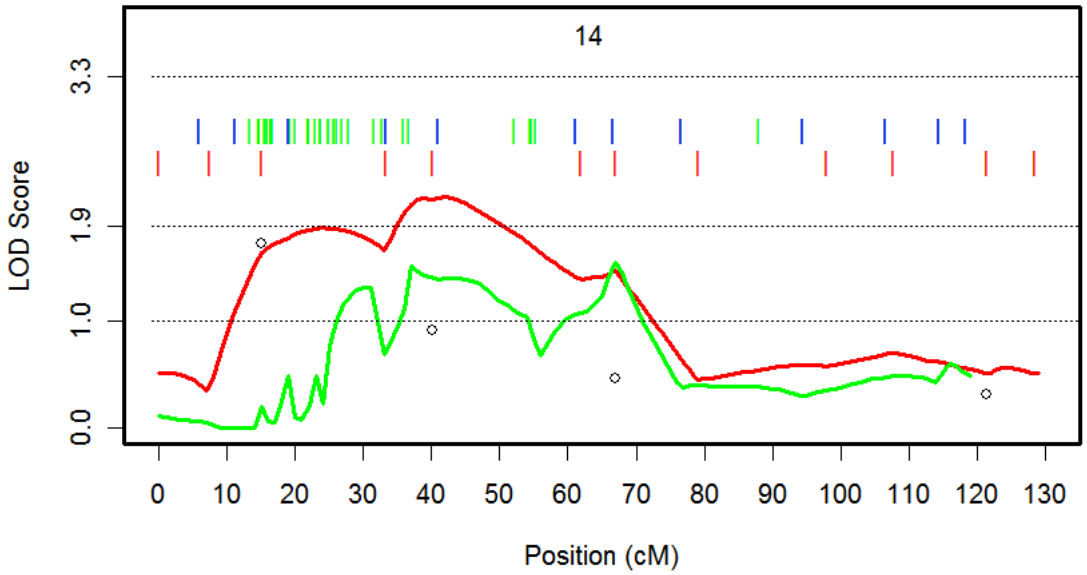

Figure 2.3.2: Evolution of linkage signals on the q-arm of chromosome 14. Single-point parametric linkage analysis (narrow dominant model) = ∘. Multipoint non-parametric genome-wide linkage analysis (Solar threshold model) = red line. Multipoint non-parametric fine-mapping linkage analysis (Solar threshold model) = green line. The original (red) and fine mapping (green) markers and their positions on the ABI and intergrated (OXSTATS) map, respectively, are given by |. The positions of the original markers on the OXSTATS map are shown in blue. Both the original and additional markers were used for fine mapping purposes.

addition of fine-mapping markers the LOD score under a threshold model fell to 1.52 (broad model). This is not unexpected for linkage analysis of complex traits, but serves an important reminder that a number of promising LOD scores in the range of, say, 2 to 3, may be false-positive findings. The power to detect linkage depends on the sample size and the marker information. The current sample size was fixed but the marker information content on chromosome 14 was maximised by adding more markers. The reduction in the test statistic suggests that the chromosome 14 finding is a false-positive, however true linkage to this region cannot be ruled out. Replication in other studies or the collection of more families and/or a population-wide association study will ultimately determine if chromosome 14 harbours one or more susceptibility genes for MDD and comorbid unexplained swelling.

For parametric linkage analysis the population frequency of each disease and marker allele is required for the calculation of the likelihood (Terwilliger and Ott, 1994). When the genotype at a locus is known for each individual in the pedigree, the specified allele frequencies at the locus have no effect on the LOD score. However, allele frequencies are important when there are founders of unknown genotype, and at loci where this is the case, the allele frequencies will have an effect on the LOD score. In the majority of genome-scans there are founders of unknown genotype at a given locus, and this is the case in the current study. When all alleles at a locus were assumed to be equally frequent, 93 markers gave a LOD score greater than 1.00. Only 30 markers gave a LOD score greater than one when allele frequencies were estimated from the genotyped individuals within the pedigrees. This finding is consistent with that of Ott *et al.* (1992) who found that choosing to use equal allele frequencies in a genome-wide scan for linkage biases the LOD scores towards linkage when founders of unknown genotype exist in the pedigree. Therefore, more attention should be paid to the

LOD scores from the analysis with estimated allele frequencies. This analysis still has limitations because allele frequencies were estimated from families selected for depression and comorbid unexplained swelling and these may not provide a true reflection of the true population allele frequencies.

The robustness of the parametric linkage analysis with regard to the estimated model parameters was checked by providing multiple models for the analysis of chromosomes 14 (the chromosome containing the maximum LOD score) and chromosome 22 (the smallest chromosome). The alternative models provided different penetrance, phenocopy rates and gene frequencies. Whilst the actual LOD score and θ did vary, the test statistics for the markers remained consistent relative to one another. That is to say that the marker which provided the maximum LOD score on a given chromosome consistently did so, and only the size of the test statistic differed. This result is encouraging, and suggests that the markers identified through the parametric linkage analysis are not dependent on the given model.

Linkage analysis was carried out using a parametric method so that the age-specific penetrances of MDD could be modelled in unaffected individuals. Furthermore, after viewing the disease transmission pattern in the four families, it was assumed that a single variant of major effect was segregating in the four families. In an attempt to increase power, most modern QTL linkage analyses use multipoint methods. Multipoint methods use the genotype information of all available markers simultaneously. The result is that multipoint methods increase the information content at a given locus and thus increase the power to detect linkage. In parametric linkage analysis the parameter which is estimated is the recombination fraction (θ), and any model misspecifications lead to an inflated θ being reported. With marker-specific linkage analysis we are interested in the

LOD score at the marker, which remains unbiased in the presence of model mispecifications provided dominance is correctly specified (Clerget-Darpoux *et al.*, 1986; Risch and Giuffra, 1992). However, incorrectly specified parameters in multipoint linkage analysis can lead to biased LOD score estimates because $\hat{\theta}$ is restricted by the marker map. For example, assume that the true $\theta = 0.1$ but, due to model misspecifications, single-point linkage analysis gives a $\hat{\theta}$ of 0.3. If, using the same parametric model, multipoint parametric linkage analysis was carried using markers spaced every 10cM ($\theta = 0.1$) then the maximum recombination fraction between the marker and disease is 0.1. Therefore, the LOD score at the marker position with a single-point $\hat{\theta}$ of 0.3 will be lower in the multi-point analysis than in the single-point analysis. Due to the problems with defining accurate genetic models for complex diseases such as MDD, and therefore the likelihood of biasing the LOD scores, a multipoint parametric analysis was not carried out.

The aetiological and genetic link between depression and unexplained swelling remains unknown. Whilst the large number of affected individuals within families and the estimated heritabilities suggests a genetic cause, this remains uncertain. The comorbidity of the diseases appears strong, with 59% of affected individuals having both depression and unexplained swelling. Assuming that genetic loci are involved in the aetiology of the disorder, the nature of these and their effects on each individual phenotype remains unclear. Genes may have a primary effect on only one of the phenotypes, with the second being a consequence of the primary condition. It is a possibility that underlying genetic mutations have an effect on a physiological pathway that leads to both depression and unexplained swelling. Alternatively, the causative genetic loci could be in linkage disequilibrium with each other and have effects on independent pathways leading to each separate disorder.

In conclusion, it is unlikely that a gene of major effect is segregating within these four families with depression and comorbid unexplained swelling. Marker D8S260 on chromosome 8q produced the maximum LOD score of 2.02 under a narrow-recessive model. Modelling the age-specific penetrance in the unaffected individuals increased the LOD scores but did not appear to bias the results.

Chapter 3

A simple grouped linear regression method for linkage analysis of censored traits

3.1 Introduction

Domestic animals and experimental species provide a unique resource for the understanding of quantitative genetic variation. Quantitative trait analysis of experimental crosses has provided many important insights into the genetics of complex traits (reviewed in Andersson and Georges (2004) and Morgante and Salamini (2003)). Several genes underlying quantitative genetic variation have been identified in the fields of animal and crop science, many of which have significant commercial potential (e.g. Jeon *et al.*, 1999; Nezer *et al.*, 1999; Frary *et al.*, 2000; Fridman *et al.*, 2000; Grisart *et al.*, 2002).

Most current quantitative trait loci (QTL) mapping techniques utilize an interval mapping approach first put forward by Lander and Botstein (1989). The approach places a hypothetical trait locus at fixed incremental positions (for example, every 1-2cM) along a map of known marker positions and tests for its effect on the trait using information from flanking markers. For a given location the basic linear

model is

$$\mathrm{y}_{ij} = \mathrm{m}_j + \mathrm{e}_{ij}, \tag{3.1.1}$$

where y_{ij} is the trait value for individual i with genotype j, mj is the mean effect of genotype j, and e_{ij} is random error ($e_{ij} \sim N(0,\ \sigma_e^2)$). The genotype of an individual at the position being tested is rarely known so the probability of an individual being each of the possible genotypes is calculated from the available marker information. Lander and Botstein (1989) implement their method using a maximum likelihood approach. The maximum likelihood method takes into account heterogeneous variances within marker classes to estimate genotype probabilities. The model parameters are estimated under both the null (no QTL) and alternative (with QTL) hypotheses. An advantage of maximum likelihood is that it uses all of the available observations on marker genotypes and trait values. The disadvantage of maximum likelihood is that it is computationally intensive and usually requires specialized software.

An alternative method, least squares regression, uses expected genotype probabilities calculated from flanking markers rather than the more complex approximation via maximum likelihood (Haley and Knott, 1992). For this approach least squares linear regression is used to estimate the effect of genotype on the trait of interest at each test position along the genome. The asymptotic equivalence of least-squares regression with maximum likelihood interval mapping has been shown through simulation (Haley and Knott, 1992) and by theoretical calculations of power (Rebai *et al.*, 1995). The least squares approach has been shown to be robust to deviations from normality in all but the most extreme situations (Visscher *et al.*, 1996; Rebai, 1997). Kao (2000) and Knott (2005) review the differences between maximum likelihood and regression QTL mapping methods.

Time-to-event traits are often nonnormally distributed and show a right-skewed distribution of trait values across all individuals. Additionally, time-dependent traits often include censored observations, which occur when the true time of the event is unknown. End-of-study censoring arises when the event of interest has not occurred by the end of the study. Within-study censoring arises if an individual is lost to follow-up during the course of the study. The loss of information due to censoring results in lower statistical power, where the greater the proportion of censoring the lower the statistical power. Some power can be recovered by modeling censored individuals in the statistical analysis; however, standard QTL mapping techniques typically do not account for this.

The field of survival analysis utilizes special methods to better use the information provided by censored observations and account for the non-normal distribution of the trait values. Traditionally, proportional hazard regression models are used to model survival traits. These methods assume that if there are two individuals, a and b, with p time-independent covariate values in vectors $\boldsymbol{Z_a}$ and $\boldsymbol{Z_b}$, respectively, the ratio of their hazards is given by

$$\frac{h(t|\boldsymbol{Z_a})}{h(t|\boldsymbol{Z_b})} = \frac{h_0(t)\exp\left[\sum_{k=1}^{p}\beta_k Z_{ak}\right]}{h_0(t)\exp\left[\sum_{k=1}^{p}\beta_k Z_{bk}\right]} = \exp\left[\sum_{k=1}^{p}\beta_k(Z_{ak}-Z_{bk})\right], \qquad (3.1.2)$$

where $h(t|\boldsymbol{Z_i})$ is the hazard for individual i at time point t, $h_0(t)$ is the baseline hazard function, and β_i is the coefficient for the effect of the ith covariate. As time dependence is only included in the baseline hazard, the ratio of the hazards of two individuals at any time point is a constant and therefore the hazards are proportional (Klein and Moeschberger, 1999). Cox (1972) proposed a semiparametric proportional hazards model that can be used to model survival data without pre-specifying the distribution of the baseline hazard. This method is widely used and has been shown to be both robust and powerful.

Parametric proportional hazard models also exist, which assume survival times follow a given distribution (for example, Weibull). Under the correct baseline hazard distribution, parametric models are more powerful than equivalent nonparametric or semiparametric methods. However, when using real data, the true underlying distribution of the baseline hazard is unknown. For this reason, Cox proportional hazards regression remains the method of choice for most survival analyses. Moreno *et al.* (2005) compared the Weibull and Cox proportional hazards models to a more conventional QTL-mapping method that ignored the nature of the survival data and found that when analyzing survival trait data the proportional hazards models have greater power.

A drawback of both proportional hazards methods is that they are computationally intensive for complex models. Models with many covariates, some of which may be time dependent, can take extensive periods of time to analyze. Several computationally intensive approaches have been proposed for QTL mapping of survival traits in line crosses (Symons *et al.*, 2002; Diao *et al.*, 2004; Diao and Lin, 2005) and outbred populations (Epstein *et al.*, 2003; Pankratz *et al.*, 2005). All of these methods are yet to be incorporated into general, widely used genome-analysis packages.

Here, a novel grouped linear regression method for the analysis of survival data that is computationally simple, robust and can be implemented in standard statistical packages is described. The method is compared to the classical Cox and Weibull proportional hazards approaches and a standard linear regression method that ignores censoring status. The relative power and robustness of the method is demonstrated through simulation and the advantages of this simplified method when compared to those currently available are discussed.

3.2 Methods

The Cox proportional hazards model has been widely adopted as the method of choice for survival analyses. However, when analyzing survival data with many tied or grouped observations, or when analyzing large datasets, the Cox model becomes computationally intensive. Grouped survival information is defined as non-continuous survival time data. Prentice and Gloeckler (1978) extended the popular Cox model for the analysis of grouped survival data. The method correctly models grouped survival data but still remains computationally intensive for large datasets. Here, a grouped approximation for continuous survival data is proposed where failure times are partitioned into a number of time periods and a linear regression model is put forward for the analysis of the grouped data. The aim of this method is to simplify the analysis of continuous survival data leading to reduced computation time and an increased ability to analyze models with greater complexity. The survival of each individual through these arbitrary time periods is coded using a series of conditional survival indicator variables, similar to that used by Madgwick and Goddard (1989) to predict breeding values in dairy cattle using lactation period survival data. Rather than adopt the maximum likelihood approach of Prentice and Gloeckler (1978) for parameter coefficient estimation a computationally efficient and robust linear regression method is put forward. The simplicity and efficiency of the model should allow the analysis to be carried out quickly on large datasets using standard statistical packages.

3.2.1 Grouped linear regression method

Survival times are sorted in chronological order, regardless of censoring status or genotype, and separated into a predefined number of groups or time-periods. The survival record for individual i during time period j is given by x_{ij}. If individual i survives interval j, then the corresponding survival record is $x_{ij} =$

Table 3.2.1: Example of the grouped linear regression group coding algorithm for two time periods

Individual	Survival	Survival record(s)
1	Censored during the first time period	NA
2	Event occurs during the first time period	1
3	Survives the first time period, the event occurs in the second	0 1
4	Survives the first time period, then is censored during the second	0

0. If individual i experiences the event during interval j, then the survival record is $x_{ij} = 1$ and there are no further survival records for the remaining intervals. If an individual is censored during a particular interval then that individual has no survival record for the current or subsequent intervals (Table 3.2.1). For any given time period, the survival record represents the conditional probability that an individual survives the current time period given that the individual survived to the start of that time period. For any individual the survival records for each group are therefore independent observations. The linear model used in the regression analysis is given by

$$x_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k t_{ik} + \beta_g g_i + \epsilon_{ij}, \tag{3.2.1}$$

where x_{ij} is the survival record for individual i during time period j, the β terms are the estimated regression coefficients, t_{ik} is an indicator variable for time period k that takes a value of 1 in time period j and 0 otherwise, g_i is the genotype for individual i, and ϵ_{ij} is the random error for individual i during time period j. The terms model the baseline hazard and $\beta_g g_i$ estimates the genotypic

effect on the hazard. Only p-1 coefficients of t_i can be estimated, as there are no non-censored individuals surviving the last time period. The significance of the effect of genotype is estimated using standard regression methodology. The resulting F-statistic (with 1 and n-p-1 degrees of freedom, where n is the number of individuals) is transformed to an approximate likelihood ratio test statistic (LRT) using the formula provided by Baret *et al.* (1998),

$$\text{LRT} = n \times log_e \left(1 + \left(\frac{1}{n-p-1}\right) F\right), \tag{3.2.2}$$

to allow comparisons with the maximum likelihood test statistic of the proportional hazards methods.

3.2.2 Simulation of data

Extensive simulations were carried out using the statistical software package R (R Development Core Team, 2005). Genotypic data were simulated at a single locus for individuals from a backcross between two fully inbred lines. Marker data were generated at a single QTL locus with possible alleles q and Q, thus assigning an individual the genotypes qq or Qq with equal probability. Phenotypic data were simulated assuming a fully penetrant QTL at the marker locus. Phenotypic data were drawn from a number of distributions (Table 3.2.2 and Figure 3.2.1). The probability density function of a Weibull distribution is given by

$$f(t) = \frac{\rho}{\lambda}\left(\frac{t}{\lambda}\right)^{\rho-1} \exp^{-\left(\frac{t}{\lambda}\right)^{\rho}}, \tag{3.2.3}$$

where ρ and λ are the shape and scale parameters, respectively. The probability density of an exponential distribution, which is a special case of a Weibull distribution where $\rho = 1$, is given by

$$f(t) = \lambda \exp^{-(\lambda)t}. \tag{3.2.4}$$

Table 3.2.2: Distributions used to simulate data

		Genotype qq		Genotype Qq	
Model	Distribution	Shape (ρ)	Scale (λ_{qq})	Shape (ρ)	Scale (λ_{Qq})
1	Weibull	2.00	10.00	2.00	9.05
2	Exponential	-	10.00	-	9.16
3	Gamma	3.65	2.19	3.65	2.43
4	Gamma	0.50	10.00	0.50	7.50

Finally, the probability density function of a Gamma distribution is given by

$$f(t) = \frac{1}{\lambda^{\rho}\Gamma(\rho)} \times t^{\rho-1} \exp^{-\left(\frac{t}{\lambda}\right)}, \tag{3.2.5}$$

where $\Gamma()$ is the gamma function.

The baseline hazard function for data drawn from a Weibull distribution is given by

$$h_0(t) = \frac{\rho}{\lambda}\left(\frac{t}{\lambda}\right)^{\rho-1}, \tag{3.2.6}$$

Thus, the ratio of the hazards for genotypes qq and Qq is

$$\frac{h(t|Qq)}{h(t|qq)} = \left(\frac{\lambda_{qq}}{\lambda_{Qq}}\right)^{\rho}, \tag{3.2.7}$$

which is independent of time (t) and thus satisfies the assumptions of a proportional hazards model. When considering model 1 (Weibull), individuals with genotype Qq are at an increased risk of approximately 22% when compared to the risk for genotype qq (for $\rho = 2$ and $\rho_{qq} = 10$, $\rho_{Qq} = 9.05$). The mean of a Weibull distribution is given by $\lambda \times \Gamma(1+\rho^{-1})$, where $\Gamma()$ is the gamma function. Thus, the simulated effect of the genotype on the hazard is equivalent to a mean

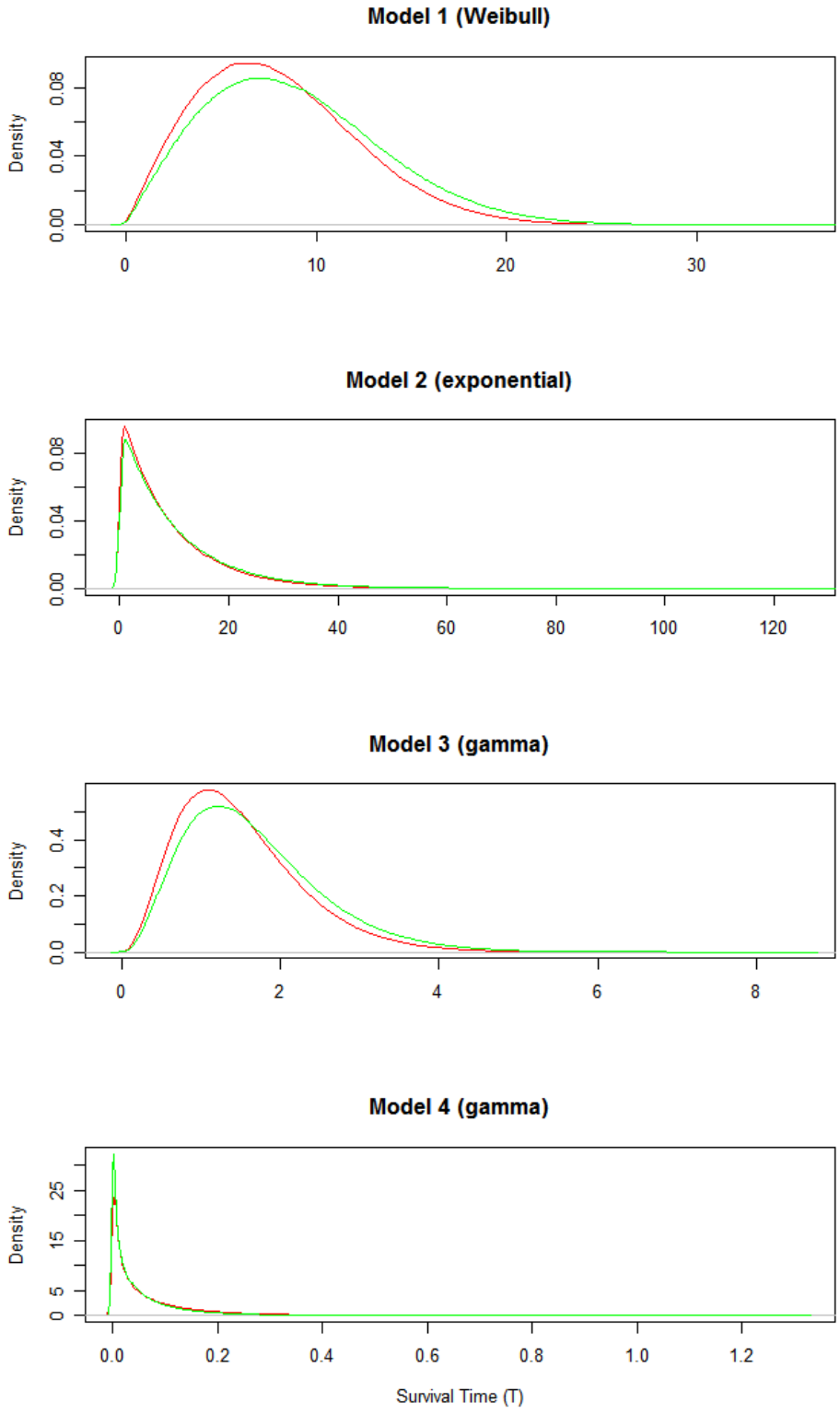

Figure 3.2.1: The four distributions from which the survival times were drawn. Figures generated by drawing 1000000 observations from each distribution.

difference in survival time of (0.18 standard deviation units). The shape and scale parameters of model 2 (exponential) were calculated to give the same ratio of hazards as that of the Weibull parameters used in model 1 but from a more highly skewed distribution. The parameters of model 3 (gamma) were chosen to approximate the means and variances of the Weibull distributions used in model 1. To test the relative robustness of the methods when phenotypic data are drawn from highly skewed distributions, data were simulated from two gamma distributions where the gamma shape parameter was 0.5. Shape parameters were chosen to approximate the hazard ratio used in model 1. The proportion of censored individuals, P_c, was varied across simulations. For the generation of censored observations the method of (Diao *et al.*, 2004) and (Diao and Lin, 2005) was implemented. Let T_i be the survival time of the ith individual, C_i is the censoring time and $I(C_i)$ is an indicator variable giving the censoring status ($0 =$ censored, $1 =$ uncensored) for individual i. Censoring times (C_i) were drawn from a uniform distribution between $0 < x \leq 1$ and multiplied by a constant τ. Diao *et al.* (2004) and Diao and Lin (2005) use a trial-and-error approach to obtain τ. Here, the value of τ was calculated via numerical integration to provide a given proportion of censored observations (see appendix 1). If C_i was less than T_i then the individual was classified as censored ($I(C_i) = 0$) and the censoring time was entered into subsequent analyses. Individuals with greater survival times are more likely to be censored. The censoring method creates both "within-study" and "end-of-study" of study censoring.

3.2.3 Analysis of simulated data

Two methods were used to group the observations (both censored and uncensored) into time periods to investigate the robustness of the grouped linear regression method to the grouping mechanism. The first grouping method (A) groups the individuals into k groups such that an equal number of observations

(either censored or uncensored) occur in each time period. The standard error of the group means is approximately equal when using this method. The second grouping method (B) groups the individuals into k groups such that an equal proportion of individuals, denoted by s, survive each time period. Within-group variances are approximately equal when using this method. For both methods the last time period contained all remaining individuals not previously experiencing the event of interest. For example, with 1000 individuals and five time periods, grouping mechanism A creates five groups of 200 individuals, while grouping method B with $s = 0.5$ creates groups of 500, 250, 125, 63, and 62 individuals. The change in mean test statistic in relation to the number of time periods and the proportion of individuals surviving each time period was investigated for each grouping mechanism.

3.2.4 Power comparisons

The power of the grouped linear regression method was compared to that of the Cox and Weibull proportional hazards models and the standard linear regression least squares approach (ignoring censoring status). The inclusion of the standard linear regression method allows some approximation of the power to be gained by including the censored observations correctly in the regression model. Grouping method B was used to separate the continuous survival times into groups for the grouped linear regression method. Values were chosen for the grouping parameters that approximately maximized the power of the grouped linear regression method. Methods were contrasted by comparing the mean test statistic from each mode of analysis given phenotypic data drawn from the same underlying distribution. To ensure this was an unbiased comparison of methods, 1000 simulations of 1000 individuals were carried out under the null hypothesis for each mode of analysis. Deviations from a chi-squared distribution with one degree of freedom were tested for using a one-tailed Kolomonov-Smirnov test.

Where a significant deviation from a chi-squared distribution was seen for a given phenotypic distribution, analysis methods were compared via empirical P-values. To calculate the empirical thresholds 10000 replicates of 1000 individuals were simulated under the null hypothesis. The proc.time( ) function in R was used to obtain run times for the model-fitting step of each mode of analysis.

3.2.5 Alternative censoring mechanisms

For the above simulations, only the τ censoring mechanism of Diao *et al.* (2004) and Diao and Lin (2005) was used. Two alternative censoring mechanisms were applied to further investigate the relative powers of the mapping methods and their robustness to the censoring mechanism. The first alternative censoring mechanism, random censoring, simulates censoring status ($I(C_i)$) by drawing from a Bernoulli distribution with probability 1-P_c. If $I(C_i) = 0$ (i.e. the individual is censored) then the censoring time (C_i) is given by (C_i) = xT_i, where x is a random uniform number between 0 and 1. This method censors individuals randomly, irrespective of survival time. Specifically, those individuals surviving longer are at no increased risk of being censored. The random censoring mechanism only creates "within-study" censoring. The second alternative censoring mechanism, here named Weibull censoring, draws censoring times (C_i) from a Weibull distribution with shape parameter ρ and scale parameter λ_c. The value of λ_c is calculated to provide a given proportion of censored observations (see Appendix 2). If C_i is greater than T_i then the individual is classified as uncensored ($I(C_i) = 1$). If C_i is less than T_i then the individual is classified as censored ($I(C_i) = 0$) and the censoring time is used in subsequent analyses. Unlike the random censoring method, individuals with greater survival times are more likely to be censored. The Weibull censoring mechanism creates "within-study" censoring.

Table 3.3.1: Grouping method A: grouped linear regression mean test statistic

Censoring	Number of groups (k)								
proportion (P_c)	2	3	4	5	6	7	8	9	10
0	1.94	7.70	8.24	8.76	9.42	9.28	9.55	9.32	*9.87*
0.1	1.91	6.72	7.14	8.26	8.44	8.48	8.91	*8.98*	8.80
0.5	3.13	5.22	5.04	5.67	5.68	5.49	5.83	5.80	*5.84*

Numbers in italics depict the maximum mean test statistic for the given censoring proportion.

3.3 Results

3.3.1 Grouping Method

Tables 3.3.1, 3.3.2, and 3.3.3 show the effect of altering the group survival proportion (s) and/or the number of groups (k) on the grouped linear regression mean test statistic. The shape and scale parameters of model 1 (Weibull) were used in the simulations. Some combinations of s and k are impossible with a sample size of 1000 individuals as the number of individuals in a group falls rapidly at low values of s, thus limiting the possible number of groups (k). For the range of groups simulated, the mean test statistic approximately increased with the number of time periods into which the data were grouped. This result was seen for both grouping methods. An increase in mean test statistic was seen when comparing grouping method B to grouping method A. This relationship only held if the optimal group survival proportion was determined correctly for grouping method B. However, this may not always be possible, in which case adopting a grouping method in which each group contains an equal number of individuals (A) leads only to a small reduction in mean test statistic. When the same proportion of individuals survive each group (grouping method B) the lowest possible group survival rate, given the number of time periods and sample size, gave a reasonable

Table 3.3.2: Grouping method B: grouped linear regression mean test statistic with no censoring

Group survival	Number of groups (k)								
proportion (s)	2	3	4	5	6	7	8	9	10
0.1	5.38	*7.09*	-	-	-	-	-	-	-
0.2	4.91	*8.50*	-	-	-	-	-	-	-
0.3	3.85	9.02	*9.39*	-	-	-	-	-	-
0.4	2.85	8.85	9.72	10.18	*10.24*	-	-	-	-
0.5	1.94	8.13	9.54	10.15	10.66	*10.78*	-	-	-
0.6	1.21	7.17	8.68	9.62	9.97	10.25	10.69	*10.76*	10.74
0.7	0.65	6.18	7.41	8.56	9.03	9.90	9.99	10.36	*10.43*
0.8	0.35	4.52	5.71	6.94	7.72	8.05	8.77	*9.48*	9.45
0.9	0.10	2.89	3.56	4.55	5.23	5.76	6.18	6.83	*6.93*

Italics depict the maximum mean test statistic for the given group survival proportion (s).

approximation to maximize the mean test statistic. For the power comparisons grouping method B was adopted, where an equal proportion of individuals ($s = 0.6$) survived each of the groups ($k = 10$). These parameters will not maximize the mean test statistic in all situations, but provide a good approximation to the optimal parameters in the situations tested and good practical guidelines for other studies of similar size.

3.3.2 Power Comparisons

The distribution of test statistics for all four modes of analysis was shown to be chi-squared under the null hypothesis for model 1 (Figure 3.3.1). This supports the use of the mean test statistic as an unbiased parameter for the comparison of methods. Figure 3.3.2 shows the mean test statistic for all four modes of analysis with varying percentages of censored survival times. In the absence

Table 3.3.3: Grouping method B: grouped linear regression mean test statistic with 50% censoring

Group survival	Number of groups (k)								
proportion (s)	2	3	4	5	6	7	8	9	10
0.1	3.46	*3.99*	-	-	-	-	-	-	-
0.2	3.23	*5.04*	-	-	-	-	-	-	-
0.3	2.66	5.06	*5.29*	-	-	-	-	-	-
0.4	2.16	4.81	5.42	5.34	*5.62*	-	-	-	-
0.5	1.78	4.72	5.04	5.45	5.69	*5.97*	-	-	-
0.6	1.61	3.97	4.65	5.23	5.23	5.69	5.62	5.63	*5.81*
0.7	1.29	3.26	3.93	4.37	5.13	5.32	5.48	*5.83*	5.76
0.8	1.10	2.34	3.23	3.68	3.91	4.32	4.65	5.13	*5.28*
0.9	1.07	1.51	1.93	2.38	3.61	2.79	3.09	3.39	*3.80*

Italics depict the maximum mean test statistic for the given group survival proportion (s).

of censoring the four analysis method have approximately equal power. The standard linear regression model has a slightly reduced mean test statistic when compared to the survival analysis methods. This difference is likely due to the non-normal distribution of the survival times. As the percentage of censored observations is increased, the standard linear regression mean test statistic falls rapidly. In comparison, that of the grouped linear regression and Cox and Weibull proportional hazard models decreases at a slower and comparable rate. The relative and actual run-times from the model fitting step of each analysis method are shown in Table 3.3.4. The procedure time for fitting 100 models showed the grouped linear regression method to be almost five times quicker than the Cox proportional hazards method and over twelve times quicker than the Weibull proportional hazards method.

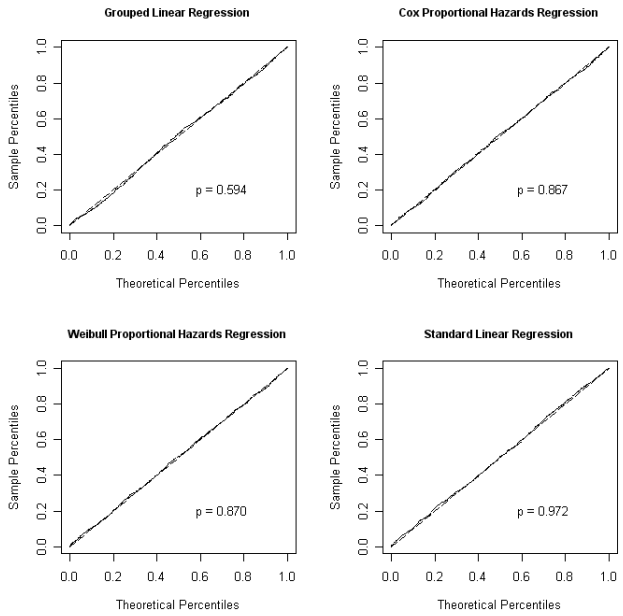

Figure 3.3.1: Q-Q plots for phenotypes simulated from model 1 (Weibull) under the null hypothesis. Dashed line denotes the 1:1 relationship between sample and theoretical percentiles. *P*-values from a one-tailed Kolomonov-Smirnov test.

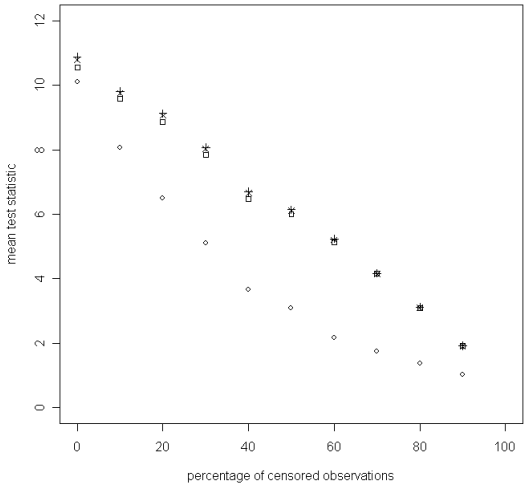

Figure 3.3.2: Mean test statistics for phenotypes simulated from model 1 (Weibull) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportion hazards regression; +, Weibull proportional hzards regression; ◇, standard linear regression.

Table 3.3.4: Time taken to fit 100 models using R (v.2.1.1)

	Relative procedure time	Actual procedure time (sec)
Standard linear regression	1	0.22
Cox proportional hazards regression	7.2	1.59
Weibull proportional hazards regression	19	4.18
Grouped linear regression	1.5	0.33

[a]Number of groups (k) = 10, group survival proportion (s) = 0.6

When simulating data under the null hypothesis from model 2 (exponential) no significant deviations were detected from a chi-squared distribution with one degree of freedom (Figure 3.3.3). When comparing the relative mean test statistics from model 2 (exponential), similar relationships to those under model 1 (Weibull) were observed (Figure 3.3.4). Given that an exponential distribution is a special case of a Weibull distribution, when ρ=1, it is not surprising to find that the parametric Weibull proportional hazards method is of equal power when compared to the Cox proportional hazards model.

When phenotypic data were simulated under the null hypothesis using model 3 (gamma), a significant deviation from a chi-squared distribution (p = 0.012) was seen when using the Weibull proportional hazards model (Figure 3.3.5). This result was not unexpected as the Weibull proportional hazards model fits a Weibull distribution to the distribution of survival times, which in this case follow a gamma distribution. To make unbiased comparisons between the four analysis methods empirical P-values were calculated for each method of analysis (Figure 3.3.6). When the phenotypes contained no censored observations the best-performing model was standard linear regression. This is not surprising as a gamma distribution with a shape (ρ) of 3.65 is not highly skewed. The

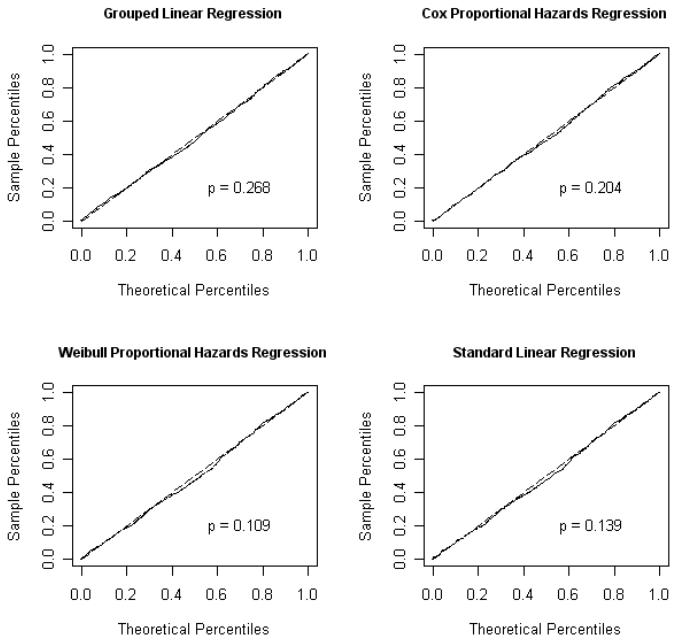

Figure 3.3.3: Q-Q plots for phenotypes simulated from model 2 (exponential) under the null hypothesis. Dashed line denotes the 1:1 relationship between sample and theoretical percentiles. *P*-values from a one-tailed Kolomonov-Smirnov test.

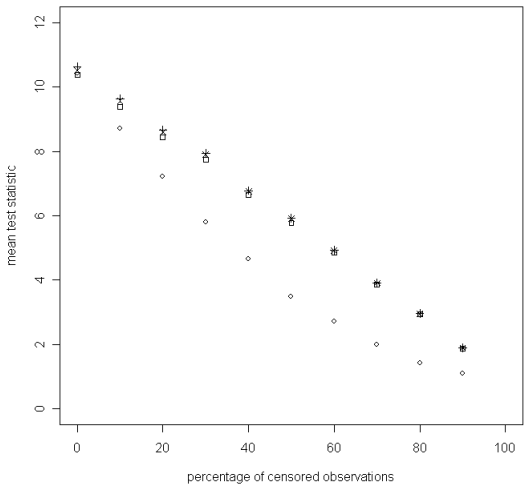

Figure 3.3.4: Mean test statistics for phenotypes simulated from model 2 (exponential) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportion hazards regression; +, Weibull proportional hzards regression; ◇, standard linear regression.

worst-performing model under no censoring was the Weibull proportional hazards model. This is most likely due to the incorrect parameterization of the baseline hazard. As the proportion of censoring is increased the three survival analysis methods outperform the standard linear regression method. The grouped linear regression method shows the same power as the Cox proportional hazards method.

The gamma distributions used for model 3 were not highly skewed. However, under the highly skewed gamma distributions of model 4 a similar pattern is seen. When analyzing data simulated under the null hypothesis the Weibull proportional hazards model again gives test statistics which are not distributed as a chi-squared distribution with one degree of freedom ($p = 0.0007$) (Figure 3.3.7). With no censoring in the sample simulated under the alternative hypothesis the Weibull proportional hazards method is the least powerful mode of analysis (Figure 3.3.8). Again, the grouped linear regression approach is of equal power to the Cox and Weibull proportional hazards models, regardless of the censoring proportion. When simulated phenotypes include censored observations, the least powerful method was the standard linear regression method.

To further check the robustness of the method, two alternative censoring mechanisms were simulated. Both the random and Weibull censoring methods only simulated "within-study" censoring. No differences in the relative powers of the methods were observed (Figures 3.3.9, 3.3.10). When censored observations were simulated using the random censoring mechanism the relative power of the four analysis methods changes when 70% or more of the observations are censored. Given that few studies attempt to analyse datasets with more than 70% censored observations this result seems insignificant.

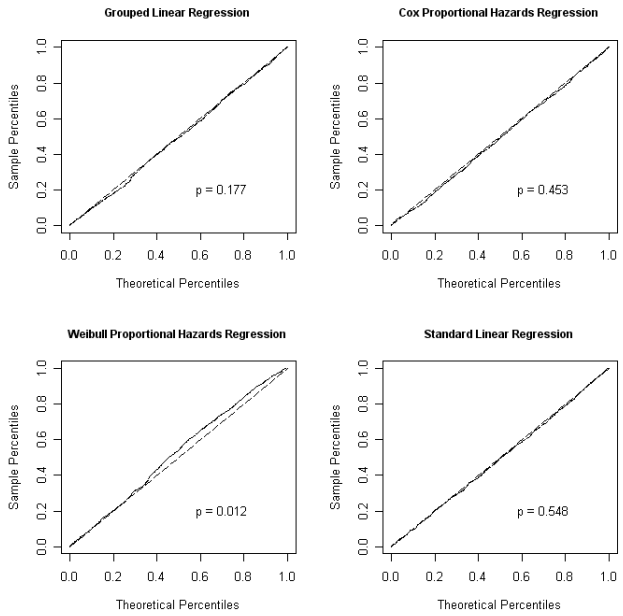

Figure 3.3.5: Q-Q plots for phenotypes simulated from model 3 (gamma) under the null hypothesis. Dashed line denotes the 1:1 relationship between sample and theoretical percentiles. *P*-values from a one-tailed Kolomonov-Smirnov test.

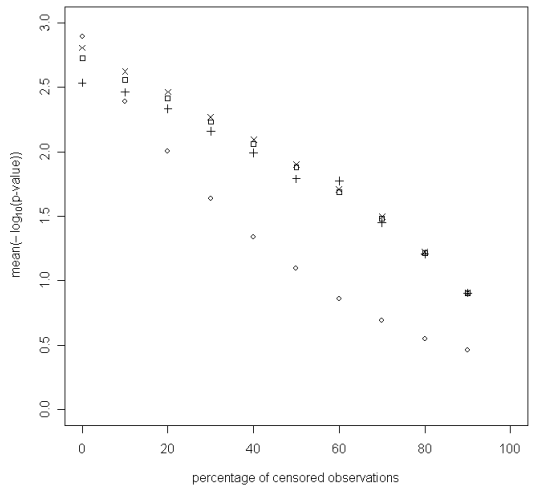

Figure 3.3.6: Mean test statistics for phenotypes simulated from model 3 (gamma) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportion hazards regression; +, Weibull proportional hzards regression; ◇, standard linear regression.

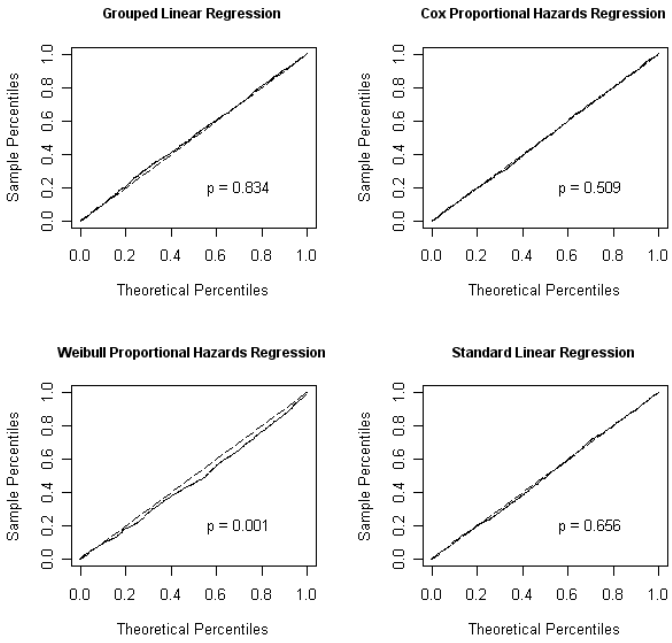

Figure 3.3.7: Q-Q plots for phenotypes simulated from model 4 (gamma) under the null hypothesis. Dashed line denotes the 1:1 relationship between sample and theoretical percentiles. *P*-values from a one-tailed Kolomonov-Smirnov test.

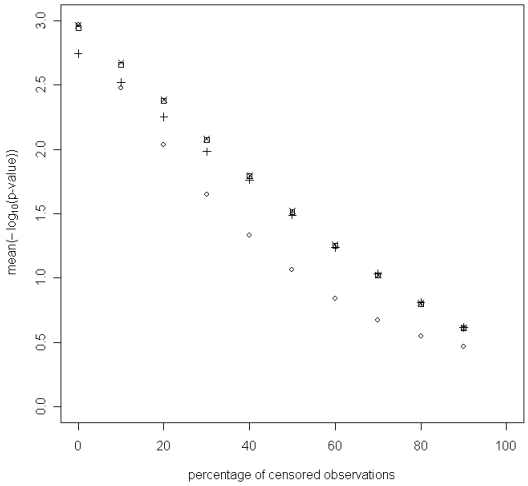

Figure 3.3.8: Mean test statistics for phenotypes simulated from model 4 (gamma) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportion hazards regression; +, Weibull proportional hzards regression; ◇, standard linear regression.

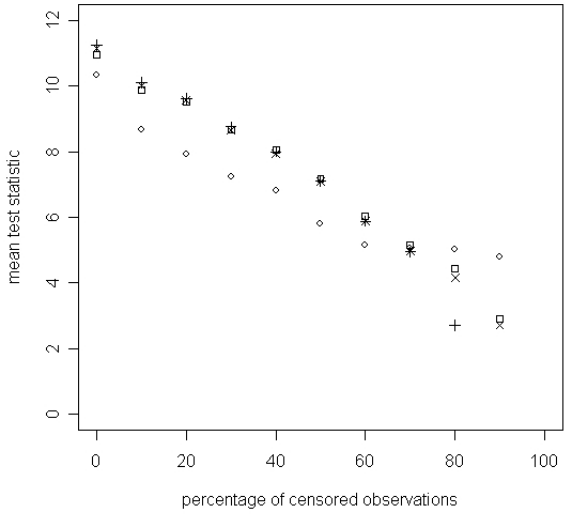

Figure 3.3.9: Mean test statistics for phenotypes simulated from model 1 (Weibull) distributions with varying proportions of random censoring. □, grouped linear regression; ×, Cox proportion hazards regression; +, Weibull proportional hazards regression; ◇, standard linear regression.

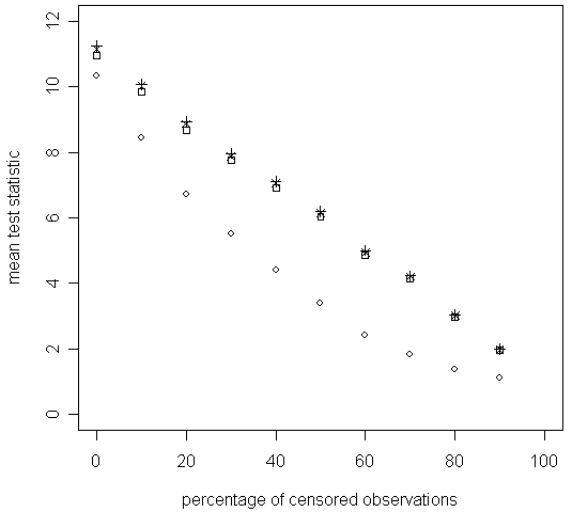

Figure 3.3.10: Mean test statistics for phenotypes simulated from model 1 (Weibull) distributions with varying proportions of Weibull censoring. □, grouped linear regression; ×, Cox proportion hazards regression; +, Weibull proportional hzards regression; ◇, standard linear regression.

3.4 Discussion

It has been shown empirically that assuming the X_ij values are Gaussian does not have a negative impact on power, at least not for the range of parameters that were considered. This is most likely because in a backcross experiment the test statistic for linkage comes from a difference in the marker means and, because of the central limit theory, these are asymptotically normally distributed (Visscher *et al.*, 1996). The assumption about the distribution of the X_ij values may be less good for extreme survival probabilities and in outbred populations in which the test statistic is based on the estimation of a variance component.

The grouped linear regression method is not only as robust and powerful as the Cox proportional hazards method but also computationally much faster. For genome-wide scans of several thousand test positions and many potential models, the saving in computation time would be considerable. The reduction in computation time would be further appreciated when carrying out permutation testing or bootstrapping. For example, consider a genome of 3000 cM with linkage analysis carried out at intervals of 1cM. If one was to carry out 1000 permutation tests on the sample then the grouped linear regression method, assuming the relative times are the same as shown in these simulations, would be 10.5 hours faster than the Cox model and 32 hours faster than the Weibull proportional hazards method. The exact savings in computation time will vary as the computation times given here for each model are estimates and will vary between software packages. The number of individuals in a study, the number of groups into which the continuous survival times are split and the proportion of tied observations in the sample are expected to have an effect on the relative run times of the various methods. Direct comparisons have been made between methods that all utilize the expected genotype probabilities at a given point when testing for a QTL at that location. Therefore, in terms of computation

time, all these models have an advantage over those models that fit a more complete, and computationally complex, maximum likelihood mixture model approach, such as the methods of Symons *et al.* (2002), Diao *et al.* (2004), or Diao and Lin (2005). Further savings in computation will be seen when comparing the grouped linear regression method to these maximum likelihood models.

Recently, Moreno *et al.* (2005) compared the power of the Weibull and Cox proportional hazards methods to standard Gaussian methods for mapping QTL in survival traits. Empirical survival times for an F2 population were sampled from a real dataset collected on salmonella resistance in mice. Data were transformed differently for each analysis method. Differing proportions of additive/dominance effects and censoring proportions were simulated. When censored observations were included in the sample, Moreno *et al.* (2005) report a significant difference in power between the proportional hazard models and standard linear regression methods for all simulation sets. This is consistent with the findings here. Moreno *et al.* (2005) also report a significant difference in power between the proportional hazards methods and standard QTL mapping methods when all observations are fully observed. This relationship was most significantly observed with an additive effect of .30 and an absence of dominance effects. Proportional hazards methods approximately provided a power of .60 while standard QTL mapping methods provided a power of .42 at the 95% level. Moreno *et al.* (2005) report that for larger additive effect sizes, and in the presence of dominance, the difference between the proportional hazards and standard QTL mapping methods disappears. In the present study when data were simulated using model 1 (Weibull), a slight reduction in the mean test statistic was observed for the standard linear regression method compared to that of the survival analysis methods. This reduction was much less marked than

that reported by Moreno *et al.* (2005). When uncensored data were simulated from model 2 (exponential) no significant differences in mean test statistic were detected when comparing the standard linear regression method to the survival analysis methods. Furthermore, when data were simulated from a gamma distribution with no censoring, the standard linear regression method was shown to have the most power at the 95% level. Reducing the effect of the QTL on the hazard to 3% (increased risk for genotype Qq when compared to that of qq) allows a more direct comparison to the simulations performed by Moreno *et al.* (2005). The reduction in QTL effect size dropped the power to detect linkage to approximately 0.6 at the 95% level, similar to the power achieved by Moreno *et al.* (2005). No significant difference was observed between the power of proportional hazards or standard methods. In addition, no differences in the power to detect linkage were seen using either a proportional hazards framework or a standard QTL mapping procedure when simulating from Weibull distributions with shape $\rho = 4$ or $\rho = 6$. Due to the way in which Moreno *et al.* (2005) simulated and transformed phenotypic data it is difficult to further examine the difference between the two study findings.

The grouped linear regression method uses binary survival indicators similar to those used by Madgwick and Goddard (1989). A grouped approach was natural for their dataset as it consisted of survival through a series of different lactation periods, which are pre-defined, biologically relevant time periods. A similar method was adopted by Meuwissen *et al.* (2002) to estimate breeding values for functional survival, in a simulated dairy cattle dataset. The authors compared both a linear and logistic regression method to a proportional hazards model and found no significant difference in estimated breeding value. In this study, a grouped linear regression method has been developed for survival traits with a continuous distribution. It has been shown that if a sufficient number

of time periods are chosen then little is lost in the way of power by grouping the data. The conditional survival probabilities (group survival indicators) are directly related to the hazard for a particular interval. Asymptotically, with many time intervals and a large number of observations per group, the conditional survival probabilities, scaled by the probability of survival until that time, are simply discrete versions of the continuous hazard. Just as the hazard function is a continuous approximation of a discrete observation (survival or death at a particular point in time), so the grouped approximation is a discrete approximation of a continuous distribution. With the grouped linear regression model the effects on the hazard are additive, whereas the usual assumption of proportional hazard models is that the effects act in a multiplicative manner. If the intervals are chosen such that the conditional survival probabilities in different time intervals are the same then the multiplicative and additive models converge.

The power of the grouped linear regression method was maximized, via simulation, prior to the comparative power analysis. However, the gain in power achieved by this is small. Most non-extreme values of s and k closely approximate the power provided by the optimal values. While it would be possible to carry out this optimization step before analyzing real data, it could be of relatively little benefit and time consuming. Thus, the robustness of the grouped linear regression method to the chosen number of time periods and group survival probability is encouraging.

Current mapping methods require specialized software for genome-wide linkage methods. The grouped linear regression method uses standard linear regression methodology and thus can be implemented in many of the widely available statistical packages, including the freely available R that was used here.

Expanding the grouped linear regression method to a genomewide level is straightforward. The ability to analyze genomewide marker data for linkage in freely available and easy to use packages is significant, especially if this can be done with little or no sacrifice in power.

In this study a backcross population was simulated; however, the extension of the method for other line crosses is relatively simple. Furthermore, it should be possible to extend the method to more complex situations such as the mapping of QTL with potentially censored data in outbred populations. Current methods for the mapping of QTL using censored data in general outbred populations are limited. Unlike in inbred lines where the QTL effect is fixed in both populations, not all individuals in an outbred population will segregate a given QTL. Furthermore, unlike fully inbred lines, each individual has a different genetic background effect. A random-effects QTL model based upon multiple 0/1 indicator variables would naturally fit into a linear mixed model framework and would allow QTL analyses in general pedigrees when a proportion of observations are censored.

In summary, a computationally efficient and fast method for the analysis of continuous survival data has been suggested. The grouped regression method is of equal power when compared to other available methods and is robust to changes in censoring proportion and mechanism and to the underlying distribution of the phenotype.

3.5 Appendix 1: Tau Censoring

Here, a value for the parameter τ used in Tau censoring mechanism is derived. For a general distribution of times to events, $f(T)$, a value of τ is required such

that a random uniform variable between 0 and τ is less than a random value for T with a probability P_c. The general solution to this problam is

$$P(T > C) = E_T(P(T > C|C)) = \int_0^\infty f(T)P(T > C|C)dT = P_c \tag{3.5.1}$$

where C is the random censoring time. The probability $P(T > C|C)$ is calculated noting that all values of T above τ are censored. Thus,

$$P(T > C) = \begin{cases} \int_0^T \frac{1}{\tau} dC = \frac{T}{\tau}; & T < \tau \\ 1; & T \geq \tau. \end{cases} \tag{3.5.2}$$

It follows that

$$P(T > C) = \int_0^\tau f(T)TdT + \int_\tau^\infty f(T)dT = 1 - \int_0^\tau f(T)dT + \frac{1}{\tau}Tf(T)dT = P_c \tag{3.5.3}$$

For the cases of the Weibull and gamma examined in this chapter, the value of the τ is solved by numerical integration.

3.6 Appendix 2: Weibull Censoring

Here, a value for the scale parameter, λ_c, used in the Weibull censoring mechanism is derived. Firstly, consider the case where no QTL effect is present. The probability density function of the time to event is

It follows that

$$f(T) = \frac{\rho}{\lambda}\left(\frac{T}{\lambda}\right)^{\rho-1} \exp^{-\left(\frac{T}{\lambda}\right)^\rho} \tag{3.6.1}$$

The value for the parameter λ_c is required from the ditribution of censoring times

$$g(C) = \frac{\rho}{\lambda_c}\left(\frac{C}{\lambda_c}\right)^{\rho-1} \exp^{-\left(\frac{C}{\lambda_c}\right)^\rho} \tag{3.6.2}$$

such that the probability that a random value of C is less than a random value of T at a pre-defined probability P_c (i.e. an observation will be censored with probability P_c).

Thus, solve

$$P(T > C) = E_T(P(T > C|C)) = \int_0^\infty f(T) \int_0^T g(C) dC dT = P_c \tag{3.6.3}$$

for λ_c. This calculation can be simplified by noting that if X is distributed as a Weibull variable with parameters ρ and λ, then X^ρ is exponentially distributed with parameter λ^ρ. As this power transformation is one-to-one, the required integral reduces to

$$\begin{aligned}
P(T > C) &= P\left(T^\rho > C^{|rho}\right) \\
&= \int_0^\infty \frac{1}{\lambda^\rho} \exp^{-\left(\frac{T}{\lambda}\right)^\rho} \int_0^\rho \frac{1}{\lambda_c^\rho} \exp^{-\left(\frac{C}{\lambda_c}\right)^\rho} dC_\rho dT_\rho \\
&= \int_0^\infty \left(\frac{1}{\lambda^\rho} \exp^{-\left(\frac{1}{\lambda^\rho} + \frac{1}{\lambda_c^\rho}\right) T^\rho} - \frac{1}{\lambda^\rho} \exp^{-\left(\frac{T}{\lambda}\right)^\rho} \right) dT\rho \\
&= -\frac{\frac{1}{\lambda^\rho}}{\frac{1}{\lambda^\rho} + \frac{1}{\lambda_c^\rho}} + 1 \\
&= \frac{\lambda^\rho}{\lambda^\rho + \lambda_c^\rho} = P_c,
\end{aligned} \tag{3.6.4}$$

which is readily solved for λ_c.

When a QTL effect is present, the probability distribution of the time to event becomes a 50:50 mixture of the distributions for the two genotypes,

$$f(T) = \frac{\frac{\rho}{\lambda_{qq}} \left(\frac{T}{\lambda_{qq}}\right)^{\rho-1} \exp^{-\left(\frac{T}{\lambda_{qq}}\right)^\rho}}{2} + \frac{\frac{\rho}{\lambda_{Qq}} \left(\frac{T}{\lambda_{Qq}}\right)^{\rho-1} \exp^{-\left(\frac{T}{\lambda_{Qq}}\right)^\rho}}{2}. \tag{3.6.5}$$

In this case, the relevant equation for P_c is

$$\begin{aligned} P(T > C) &= P\left(T^{\rho} > C^{|rho}\right) \\ &= \frac{\lambda_{qq}^{\rho}}{2\left(\lambda_{qq}^{\rho} + \lambda_{c}^{\rho}\right)} + \frac{\lambda_{Qq}^{\rho}}{2\left(\lambda_{Qq}^{\rho} + \lambda_{c}^{\rho}\right)} \\ &= P_c \end{aligned} \tag{3.6.6}$$

giving the quadratic equation in λ_c^{ρ}

$$P_c\left(\lambda_c^{\rho}\right)^2 + \left(P_c - \frac{1}{2}\right)\left(\lambda_{qq}^{\rho} + \lambda_{Qq}^{\rho}\right)\lambda_c^{\rho} + (P-1)\left(\lambda_{qq}^{\rho} \times \lambda_{Qq}^{\rho}\right) = 0 \tag{3.6.7}$$

from which the positive solution is used in generating censoring times.

Chapter 4

Estimation of variance components for age at menarche in twin families

4.1 Introduction

As discussed in the previous chapter, the semi-parametric Cox proportional hazards method remains the method of choice for most survival analyses. However, the Cox model relies on the survival times of individuals being independent, and this is not always the case. Individuals can be grouped in such a way that their survival times become correlated. For example, individuals could belong to the same family. If this were the case, and there was a common environmental effect on survival time, then one would expect the survival times of family members to be correlated. If the Cox model is applied to non-independent data then the model parameters are overestimated (Wei *et al.*, 1989). Special methods are needed to analyse data when survival times are correlated.

Frailty models have been derived to model non-independent survival data. A frailty is an unobserved random effect which acts multiplicatively on the baseline hazard. A shared frailty is a random effect which is the same for all members of a group, for example a family effect (Xue and Brookmeyer, 1996). With this model, families with a large frailty will experience the event earlier than families with a small frailty. The model therefore allows for the presence of both 'frail' and

'robust' families (Klein and Moeschberger, 1999). However, for genetic studies, fitting only a single family effect is unappealing as people within a family are related to differing degrees. To model the genetic relationship of individuals within a family a correlated frailty method must be adopted. These methods fit a per-individual random effect which is correlated according to a relationship matrix. The relationship matrix denotes the degree of genetic relatedness for all pairs of individuals within a family. It is the variance-covariance matrix of the additive genetic random effects. Each off-diagonal element is the additive genetic coefficient of relatedness for that pair, which is equal to twice the kinship coefficient (Lynch and Walsh, 1998). In a non-inbred population, the diagonals of this matrix are 1. For example, in a family of three siblings, a 3x3 matrix would show the covariance of the siblings to be 0.5 (the matrix off-diagonals) and the variance of each sibling to be 1 (the matrix diagonals). Therefore, the dependency between individuals is fully accounted for by the random effects in the model. While most random effects methods assume that the frailties follow a gamma distribution, a Gaussian random effects model is most easily generalised to arbitrary covariance matrices. Ripatti and Palmgren (2000) proposed a mixed effects Cox model which includes both fixed and random effects and is given by

$$\lambda(t) = \lambda_0(t) \exp^{\mathbf{X}\beta + \mathbf{Z}b} \tag{4.1.1}$$

where $\lambda_0(t)$ is the baseline hazard, $\mathbf{X}$ is a fixed effect matrix, $\mathbf{Z}$ is a random effect matrix and β and b are the corresponding parameter vectors.

Like the traditional Cox model, the mixed effects Cox model is semi-parametric and does not require the distribution of the baseline hazard to be specified. In addition, the model retains the proportional hazards framework as it is assumed that the conditional individual-specific hazards are proportional over time. The model can be applied to estimate variance components in outbred populations. If

an IBD matrix and a relationship matrix are included as random effects matrices, the model can also be used for mapping QTL in outbred populations. An IBD matrix is a realised version of the additive genetic relationship matrix at a given locus. It is the covariance matrix of the random QTL effects at that locus. Off-diagonal elements are half the probability that a pair shares 1 allele identical by descent (IBD) plus the probability that the pair shares 2 alleles IBD. In a non-inbred population, these elements are equivalent to the proportion of alleles shared IBD between the pair. Diagonals are equal to unity. Using this statistical model, Zhao (2005) found that marker D4S1645 contributed significantly to the variance in alcohol dependency in 143 Genetic Analysis Workshop 14 families drawn from the Collaborative Study on the Genetics of Alcoholism.

In this chapter, the age at onset trait of interest is age at menarche (AAM). Age at menarche, the time of first menstrual period, is an important developmental milestone in females and is a clearly defined event in the course of female pubertal development. Age at menarche is also thought to be an important evolutionary trait (Kirk *et al.*, 2001).

Age at menarche is a complex trait which is determined by an array of genetic and environmental factors. Several twin studies have been carried out to partition inter-individual trait variation in age at menarche into genetic and environmental components and their findings are summarised in Table 4.1.1. This is an important first step in understanding the genetics of a trait, and one which should be carried out prior to gene mapping experiments. Genetic factors clearly play a role in age at menarche, with monozygotic (MZ) twin correlations in the range of 0.51-0.95, and dizygotic (DZ) twin correlations in the range of 0.17-0.58 corresponding to heritabilities in the range 0.30-0.95. Studies which make use of other family structures support these findings. A

recent study of age at menarche, carried out using family data from the Fels Longitudinal Study (Roche, 1992), reported a heritability (h^2) of 0.49 (95% CI = 0.24-0.73) (Towne *et al.*, 2005). The study analysed data from 371 white females from extended families and found not only a significant genetic effect on age at menarche, but also a year of birth effect that explained 0.02 of the residual phenotypic variation. A recent twin study has suggested that common genes (but different environments) influence the sequence of pubertal events (van den Berg *et al.*, 2006) of which age at menarche is the most readily scored.

Of the six twin studies described in Table 4.1.1, three were carried out on adolescent samples. Kaprio *et al.* (1995) collected age at menarche data from 323 Finnish twin pairs who were within 3 months of their sixteenth birthday at the time of data collection. No censored observations were present in the sample. Loesch *et al.* (1995) analysed age at menarche data from a small sample of adolescent twin pairs from Poland. The twins were examined annually throughout adolescence, up to the age of 18. As a result of the long period of follow-up, no censored observations were present in the sample. van den Berg *et al.* (2006) recruited two cohorts of twins with average ages of 12.2 years and 12.4 years, respectively. Due to the young age of the twins, the two cohorts contained many censored age at menarche observations, with the first and second cohorts having 86% and 74% censored individuals, respectively. Neither of the Dutch cohorts were followed-up to remove the censored observations from the sample.

It is important to use age at menarche data collected from adolescent samples when carrying out genetic analysis of age at menarche because reports of age at menarche in adult samples have been shown to be inaccurate. In a sample of 60 women with a known age at menarche, Damon *et al.* (1969) reported a

Table 4.1.1: Previous twin studies of age at menarche

Study	Number of pairs	r_{MZ}	r_{DZ}	Best fitting model	Heritability (h^2)
Treloar and Martin (1990)*	MZ: 1,177 DZ: 711	0.65	0.18	ADE	0.61-0.68†
Meyer *et al.* (1991)*	MZ: 1,178 DZ: 711	0.65	0.18	ADE	0.71
Kaprio *et al.* (1995)	MZ: 234 DZ: 189	0.75	0.31	AE	0.74
Loesch *et al.* (1995)	MZ: 44 DZ: 42	0.95	0.58	-	0.95
Kirk *et al.* (2001)	MZ: 1,373 DZ: 1,310	0.51	0.17	ADE	0.50
van den Berg *et al.* (2006)	MZ: 36 DZ: 34	0.56^1	0.58^1	ACE	0.30
	MZ: 39 DZ: 14	0.74^2	0.34^2		

MZ = Monozygotic twin pair, DZ= Dizygotic twin pair. Best fitting model described in terms of (A) additive genetic, (C) common environment, (D) dominant genetic and/or (E) non-shared environmental effects. *Treloar and Martin (1990) and Meyer *et al.* (1991) studies used the same cohort of individuals. †, heritability varied between several analysed age-cohorts. van den Berg *et al.* (2006) give the twin correlations from two study cohorts, denoted here by ^{1} and ^{2}, respectively. The number of pairs given for the van den Berg *et al.* (2006) study only includes pairs where age at menarche is known for both individuals.

correlation between actual and recalled age at menarche of 0.78 when recall was requested approximately 19 years post-menarche. When recall was attempted 39 years post-menarche the correlation coefficient between actual and recalled age at menarche was 0.6 (Damon and Bajema, 1974). Koo and Rohan (1997) showed that even over a period of three years the accuracy of age at menarche recall decreases. They reported that after an interval of 1-2 years only 59% of females could recall the exact year and month of their menarche, whilst 77% of women were accurate to within one month. If, in an attempt to remove the recall bias of adult samples, one ascertains adolescent samples then a large proportion of the sample could be yet to experience menarche (van den Berg *et al.*, 2006). A study design which both uses adolescent females and has sufficient follow-up to remove censoring is therefore required for accurate studies into the genetics of age at menarche. To date, the only prospective and longitudinal twin study that has been carried out to look at the genetics of age at menarche is the Loesch *et al.* (1995) study, and this only had a sample size of 44 MZ and 42 DZ twin pairs.

4.2 Methods

In the present study, a sample of adolescent MZ and DZ pairs, and their siblings, is used to partition the variance of age at menarche into genetic and environmental components. The adolescent twins were first seen close to their 12th birthday and followed up at ages 14 and 16, so the age at menarche data should be accurate and contain only a small proportion of censored individuals. Two methods are used to partition the variance of age at menarche into underlying components: a standard method which does not statistically account for the censored observations in the data, and a correlated frailty model which does account for the censored nature of the data.

4.2.1 Adolescent twin families

Adolescent twins and their families were recruited for an ongoing study of melanoma risk factors at Queensland Institute of Medical Research, Australia. Twins were interviewed at ages 12 and 14. Non-twin siblings were asked to attend the interview if they were more than ten years old and had not previously attended (i.e. the siblings attend the clinic for interview once only). As part of the clinical protocol, described by Zhu *et al.* (1999), female adolescents were asked during interview to provide the date of their first menstrual period. Date of first menstrual period and date of birth were used to calculate the age at menarche for each individual (in months). Two age at menarche measures were potentially available for the twins who attended both the age 12 and age 14 interviews. For individuals with repeat observations the correlation between the age at menarche reported at age 12 and that given at age 14 was calculated. The present analysis uses age at menarche data collected between May 1992 and February 2006.

A second sample of adolescent twin families was recruited to an ongoing study of cognitive ability, again at Queensland Institute of Medical Research, Australia. Twins were interviewed at age 16, with siblings asked to attended if they were 10 years of age or older. As part of the clinical protocol, described previously by Wright *et al.* (2001), female participants were asked by a research nurse to provide the date of their first menstrual period. The same procedure as implemented in the melanoma risk factor study was used to calculate the age at menarche. A subset of the 16 year old cohort (324 individuals) was asked, via a telephone interview at a later date, to give their age at menarche. These individuals attended for interview before age at menarche data was introduces as part of the cognitive ability study protocol. This current analysis uses age at menarche data collected between July 1996 and February 2006.

Where two or more age at menarche estimates were available for an individual, the estimate provided at the first data collection following menarche was used. It was assumed that the recall closest in time to menarche would be the most accurate. The date of interview was recorded for all individuals. If an individual was censored (i.e. the true age at menarche was unknown because the individual had not started menstruating at the time of last interview), the age at last interview was used in the analysis.

For ease of computation, 30 females were removed from the data set because they were either the last born member of a triplet or members of a second twin pair. In total, the data consisted of age at menarche information for 1,351 adolescent twins and their siblings, 226 (16.73%) of whom had a censored age at menarche. Univariate outliers were identified as individuals reporting an age at menarche less than 104 or greater than 208 months. In total, 6 individuals were identified as outliers and removed from further study.

Siblings were asked to attend the clinic for interview if they were 10 years of age or above. Twins attended the clinic for interview at ages 12, 14 and/or 16. It is likely that a 10 to 12 year old sibling would not have started menstruating, and therefore would be censored for age at menarche. In this scenario, the age at last seen (10-12 years) was used as an age at menarche. The minimum censored age at menarche (age at last seen) for a twin is 12 years. Not only can siblings attend the clinic at a younger age than the twins, but they can also attend when older than 16 years of age. The range of the censored age at menarche estimates in the siblings and twins is 116 to 208 months and 144 to 194 months, respectively. The presence of siblings older than 16 years will have a lesser effect on the age at menarche variance because it is unlikely that these

individuals will be censored. Thus, the age at last seen was not used for these individuals. However, given the findings of Koo and Rohan (1997) regarding the reduction in age at menarche recall accuracy over a period of 1-2 years, it is likely that older siblings provided a less accurate estimate of age at menarche.

To ensure that all individuals within the study had the same opportunity to experience menarche, the siblings with an age at interview of less than 12 years were removed from the study. A total of 223 non-twin sisters, including 21 sibling pairs, had an age at interview greater than 12 years. The final sample consisted of 1,302 adolescent females, 184 (14.13%) of whom had a censored age at menarche.

4.2.2 Estimation of variance components

Twin pair correlations can be used to decompose inter-individual trait variation into genetic and environmental components. Inferences are based on the genetic similarity of monozygotic versus non-monozygotic (DZ twins and siblings) twin pairs; monozygotic twins share all their genes in common and non-MZ pairs share on average half their genes in common. When the phenotypic correlation between monozygotic twin pairs is greater than that of non-MZ twin pairs, it is assumed that genetic influences underlie the increased familiality. If the phenotypic correlation between non-MZ pairs is more than half the phenotypic correlation between monozygotic twin pairs, a common environmental effect on the trait is indicated. If the phenotypic correlation between non-MZ pairs is less than half the phenotypic correlation between monozygotic twin pairs, then this indicates genetic dominance or an epistatic (gene-gene interaction) effect on the trait. When variance components were being estimated the phenotypic mean and phenotypic variance of the whole sample was used (i.e. separate means and variance were not estimated for siblings, MZ twins and DZ twins). Furthermore, it was assumed that the MZ-DZ, MZ-sibling, DZ-DZ, DZ-sibling

and sibling-sibling covariances were equal, and a single correlation was estimated for these non-MZ pairs.

The twin analysis method described above relies on the assumption that the only difference between MZ and non-MZ pairs is the degree of genetic relationship between the individuals. However, the common environmental influence on MZ and non-MZ pairs could also be different. For example, the fact that MZ twin pairs look more alike could lead to them being treated more similarly. More effort may be made to keep MZ twins together throughout their schooling, or parents may have a greater tendency to dress them alike. The classic twin analysis does not account for a greater common environmental component in MZ twins compared to non-MZ pairs. However, results from other methods (for example, extended family studies such as that carried out by Towne *et al.* (2005) for age at menarche) have been in general agreement with the findings of twin studies.

The means and variances of the monozygotic twins, dizygotic twins and siblings were calculated. The covariance and correlation was calculated for monozygotic twin pairs, dizygotic twin pairs, sibling pairs and non-MZ pairs using MX (Neale *et al.*, 2002).

'Non-survival analysis' method: Here, the analysis is carried out assuming age at menarche is normally distributed and that censoring can be ignored. The amount of phenotypic variance explained by additive genetic (A), common environmental (C), specific environmental (E) and dominance/epistatic effects (D) was estimated through structural equation modelling, using the software package MX (Neale *et al.*, 2002). The additive genetic matrix is equivalent to the relationship matrix previously described. The common environment matrix is unity for pairs of individuals from the same family and zero otherwise. The

specific environmental matrix is zero for all pairs of individuals and unity for each individual with itself. The dominance/epistatic matrix is 0.25 for sibling pairs, unity for MZ twins and zero otherwise. Censored observations were included in the analysis by giving the age at last seen as an age at menarche. The censored nature of the data was not accounted for in the statistical analysis. To allow a direct comparison to the survival analysis method, described later, the mean and variance of age at menarche was equated across all zygosity groups. An ACE model, or if the non-MZ pair correlation was less than half the MZ correlation, the ADE model, was fitted to the age at menarche data. More simplified models were fitted in turn to test whether A, C (or D), or both parameters could be dropped from the full model. The fit of each sub-model was assessed by the difference in log likelihood between the sub and full models. Twice the difference in log likelihood follows a χ^2 distribution with the degrees of freedom equal to the difference in degrees of freedom between the sub and full models. For variance components, the distribution of likelihood ratio test statistics under the null hypothesis is a 50:50 mixture of a point mass at zero and a chi-squared distribution with one degree of freedom (Self and Liang, 1987; Stram and Lee, 1994). A chi-squared goodness of fit test was used to directly compare the full model to the reduced models. P-values were calculated from a chi-square distribution with 1 degree of freedom and subsequently divided by a factor of two. A P-value of less than 0.05 indicates a significant reduction in the fit of the model. If the ACE model provides the best fit to the data, the heritability is reported as the proportion of the variance explained by additive genetic effects. If the ADE model is the best fitting model then the broad-sense heritability is the sum of the proportions of variance explained by additive and dominant genetic effects.

Survival analysis method: To investigate the effect of modelling the censored observations correctly in the biometrical analysis, the general mixed-effects Cox

model of Ripatti and Palmgren (2000) was used. The analysis was carried out using the UNIX-based S-PLUS package KINSHIP (Therneau, 2003). The package was ported into the R environment (R Development Core Team, 2005) for ease of use and free availability. Censored observations were again included in the analysis by inputting the age at last seen as an age at menarche. A status vector was included to distinguish between censored and fully observed age at menarche data, where 0 indicated a censored observation and 1 indicated a fully observed age at menarche. The *makekinship* function within the KINSHIP package creates a symmetrical relationship matrix of form

$$\mathbf{A} = \begin{pmatrix} L & M' \\ M & N \end{pmatrix}, \tag{4.2.1}$$

where $\mathbf{A}$ is the relationship or kinship matrix, L and N are the relationships of individuals 1 and 2 with themselves (which are always 0.5), and M is the relationship of individual 1 to individual 2 (which for siblings is 0.25). M' is identical to M and contains a pointer to the $\mathbf{A}_{12}$ element of the matrix. Dummy parents were created and included to establish the familial relationships.

The *makekinship* function does not account for identical twins so the relationship matrix was manipulated manually to give the correct genetic relationship. If, in a family of three siblings, siblings 1 and 2 are identical twins then the corresponding section of the relationship matrix is given by

$$\mathbf{A} = \begin{pmatrix} 0.50 & 0.25 & 0.25 \\ 0.25 & 0.50 & 0.25 \\ 0.25 & 0.25 & 0.50 \end{pmatrix} \Rightarrow \begin{pmatrix} 0.5001 & 0.50 & 0.25 \\ 0.50 & 0.5001 & 0.25 \\ 0.25 & 0.25 & 0.5001 \end{pmatrix}, \tag{4.2.2}$$

where the matrix element $\mathbf{A}_{12}$ has been multiplied by a factor of two to account for the identical twins. So the relationship matrix remained positive definite, a small constant (0.001) was added to the diagonal of the full matrix. The **A** matrix fitted using MX is twice the relationship matrix fitted using COXME. A common-environment matrix (**C**) was created by changing all non-zero elements of the relationship matrix to 1. A small constant (0.0001) was again added to the diagonal of matrix to make the matrix positive definite. A matrix which models dominance effects was also created, where the diagonal of the matrix was fixed to 1.001, MZ pairs had a covariance of 1 and non-MZ pairs had a covariance of 0.25. The variance explained by non-shared environmental effects cannot be estimated by COXME because the error term is not in the linear predictor section of the model. The AC model, or, if the DZ correlation was less than half the MZ correlation, the AD model, was fitted to the age at menarche data. Simplified models, A only and C/D only, were then fitted in turn to test if the A or C/D parameters could be dropped from the full model. The fit of each sub-model was evaluated using the difference in integrated likelihoods between the sub and full models. The difference in integrated likelihoods follows a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the models. The chi-squared statistic was subsequently divided by a factor of two to ensure the correct distributional properties. A P-value of less than 0.05 indicated a significant reduction in the fit of the model.

The interpretation of the variance (θ) from COXME is not straightforward. As COXME fits a semi-parametric Cox model to the data, a one-to-one transformation of the age at menarche data does not change the variance explained by the model parameters. Thus, a transformation of the time scale to make the hazard constant will not affect the variance component estimates or the likelihoods of the fitted models. A constant hazard indicates that the

survival times are distributed exponentially. The exponential distribution is a special case of a Weibull distribution. Therefore, the method first put forward by Yazdi *et al.* (2002) for interpreting the variance from a parametric Weibull proportional hazards threshold model can be used. del P. Schneider *et al.* (2005) extended the method to take into account multiple random effects and to make better use of the proportion of censored observations. Using the method of del P. Schneider *et al.* (2005), the heritability (h^2) of age at menarche from an ACE model is given by

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_c^2 + \frac{1}{1-P_c}}, \tag{4.2.3}$$

where σ_a^2 is the variance explained by the additive genetic effects (the relationship matrix), σ_c^2 is the variance explained by the common environment, and P_c is the proportion of censored observations. With this interpretation, the proportion of variance explained by non-shared environmental effects is quantified by $\frac{1}{1-P_c}$ (the error term). Replacing σ_c^2 with σ_d^2 allows one to quantify the proportion of variance in an ADE model explained by dominance (or epistatic) effects.

With the survival analysis method, the covariance between the random (frailty) effects (e.g., A and C/D) are defined in the model for the hazard. That is, (co)variances act linearly on the log-hazard scale. The random effects allow the hazards of individuals to be proportional to each other. A covariance of the random effects implies that the hazards of a pair of individuals are correlated. Since there is a non-linear relationship between time-to-event data and the hazard function, the relationship between covariances on the (log)hazard scale and on the time-to-event scale will also be non-linear. A large covariance on the log-hazard scale implies more correlated time-to-event observations.

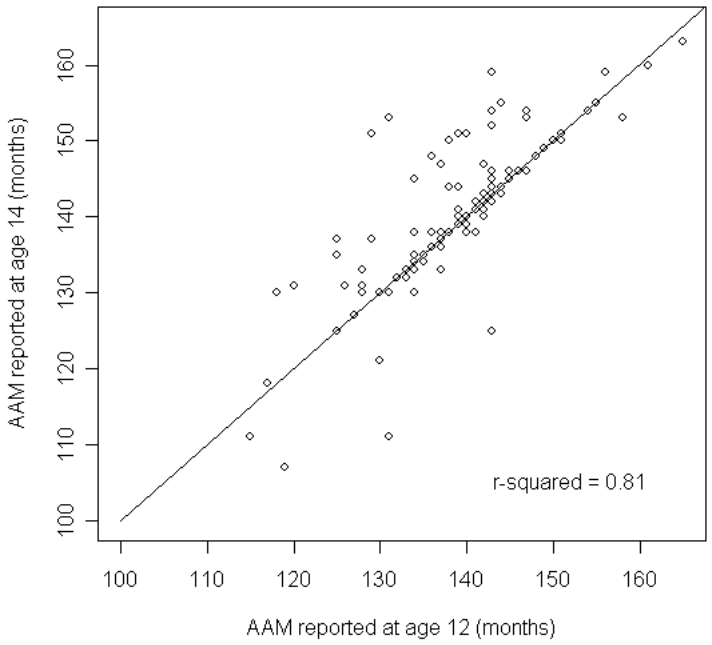

Figure 4.3.1: Age at menarche reported at age 12 versus age at menarche reported at age 14. From the sample of 1,302 individuals, 140 individuals reported an age at menarche at both age 12 and age 14.

4.3 Results

Of the total 1,302 females with age at menarche data, 140 individuals reported an age at menarche at both age 12 and age 14. The correlation between the age 12 and age 14 estimates was 0.81 (Figure 4.3.1). Summary statistics, including means and variances, are given for the MZ twins, DZ twins and siblings in Table 4.3.1. The within-pair correlations for MZ/MZ, DZ/DZ, sibling/sibling and twin/sibling pairs are given in Table 4.3.2. The MZ correlation is significantly greater than both the DZ and sibling correlation, suggesting a genetic effect on age at menarche.

The mean age at menarche within the total sample is 154.9 months, and the variance is 174.3. The sibling variance is not significantly different to that of

Table 4.3.1: Statistical descriptions of the AAM data for the MZ twins, DZ twins and their non-twin sisters

	MZ twins	DZ twins	Non-twin sisters
Number of individuals	446	633	223
Mean (months)	154.6	154.3	156.6
Variance	175.2	169.1	183.1
Number of censored individuals	62	108	14
Proportion of censored individuals	0.139	0.171	0.063

The DZ sample contains both female-female DZ pairs and female twins from opposite sex twin pairs. Only individuals with an age at last interview of 12 years or more are included.

the MZ or DZ twins. There is a large, though not significant, difference in the correlations between DZ and sibling pairs. The siblings have a much lower correlation than the DZ twins. This is perhaps not unexpected given that the sibling-sibling correlation is calculated using only 21 sister pairs. Assuming a single mean and variance for all zygosity classes, the non-MZ pair correlation for age at menarche is 0.44 (95% CI = 0.35-0.52). Under the above assumptions, the MZ pair, DZ pair and sibling pair correlations are 0.82 (95% CI = 0.78-0.85), 0.57 (95% CI = 0.46-0.66) and 0.23 (95% CI = 0.00-0.55), respectively. The non-MZ correlation is more than half the MZ correlation, therefore suggesting that an ACE model is the correct model to fit to the data.

4.3.1 'Non-survival analysis' method

The results of the non-survival biometric analysis, carried out using MX, are given in Table 4.3.3. The best fitting model is an AE model. The heritability (h^2) of age at menarche within the current adolescent sample, calculated using MX, is 0.82 (95% CI = 0.78-0.85).

Table 4.3.2: Pairwise statistics for the age at menarche data in MZ pairs, DZ pairs, sibling pairs and twin/non-twin pairs.

	MZ/MZ pairs	DZ/DZ pairs	Sib/Sib pairs	Twin/Sib pairs
Number of pairs	223	164	21	226
Correlation (r^2)	0.82	0.57	0.23	0.35
r^2 95% confidence interval	0.78-0.85	0.46-0.66	0.00-0.55	0.21-0.46

All correlations and confidence intervals were calculated using MX (assuming one mean and one variance across all zygosity groups). Only individuals with an age at last interview of 12 years or more were included.

4.3.2 Survival analysis method

The results of the biometrical analysis carried out using COXME are given in Table 4.3.4. As suggested by the MZ and DZ twin pair correlations, the best fitting model is an ACE model. The approximation of the heritability (h^2) of age at menarche within the current adolescent sample, calculated using COXME, is 0.57. COXME does not currently allow the calculation of confidence intervals.

4.4 Discussion

In the present study, if the age at menarche for an individual is censored, the age at last seen is used in the analysis. This method is perhaps the simplest method to adopt, and if the censored nature of the observations is accounted for in the statistical model it introduces no bias to the results. However, if the censored observations are not taken into account when analysing the data, as with the MX analysis carried out here, there is a potential for bias to be introduced in the estimation of the variance components. The different ascertainment of the twins and the siblings leads to a greater proportion of the siblings being censored (20.30% of all siblings regardless of age at last seen) than the twins (15.76% of

Table 4.3.3: Results from multiple models used to test alternative sources of variation in age at menarche after the removal of siblings less than 12 years old (implemented using a non-survival analysis approach with MX)

Model	V_A (95% CI)	V_C (95% CI)	V_E (95% CI)	A (95% CI)	C (95% CI)	E (95% CI)	-2LL	△-2LL	df	P
1. ACE	11.50 (10.14-12.54)	3.24 (0.00-6.31)	5.61 (5.13-6.17)	0.76 (0.58-0.85)	0.06 (0.00-0.22)	0.18 (0.15-0.22)	10,087.23	-	1,297	-
2. AE*	11.91 (11.27-12.56)		5.58 (5.11-6.12)	0.82 (0.78-0.85)		0.18 (0.15-0.22)	10,087.67	0.45	1,298	0.25
3. CE		10.18 (9.40-10.99)	8.47 (7.98-9.00)		0.59 (0.53-0.64)	0.41 (0.36-0.47)	10,171.87	84.64	1,298	0.00
4. E			13.19 (12.69-13.71)			1 (1)	10,403.22	315.33	1,299	0.00

Units of variance are months^2. Best fitting model marked by *. Alternative models were tested using the chi-squared goodness of fit test to compare the fit of the saturated model (ACE) to reduced models as potential components of variation were removed. V_A = Variance explained by additive genetic effects, V_C = Variance explained by common environmental effects, V_E = Variance explained by non-shared environmental effects. A = Proportion of variance explained by additive genetic effects (heritability), C = Proportion of variance explained by common environmental effects, E = Proportion of variance explained by non-shared environmental effects. -2LL = -2×loglikelihood, df = degrees of freedom.

Table 4.3.4: Results from multiple models used to test alternative sources of variation in age at menarche after the removal of siblings less than 12 years old (implemented using correlated frailty Cox models in COXME)

Model	V_A	V_C	A	C	E	-2ILL	△-2ILL	df	P
1. ACE*	3.42	1.42	0.57	0.23	0.20	13,534.20	-	2	
2. AE	5.15	-	0.82	-	0.18	13,539.67	5.47	1	0.010
3. CE	-	2.01	-	0.63	0.37	13,573.01	38.81	1	0.000

Best fitting model marked by *. Alternative models were tested using the chi-squared goodness of fit test to compare the fit of the saturated model (ACE) to reduced models as potential components of variation were removed. V_A = Variance explained by additive genetic effects, V_C = Variance explained by common environmental effects. A = Proportion of variance explained by additive genetic effects (heritability), C = Proportion of variance explained by common environmental effects, E = Proportion of variance explained by non-shared environmental effects, $1/1\text{-}P_c = 1.165$ (see Equation 4.2.3). All proportions of variance were calculated using Equation 4.2.3. -2ILL = -2×Integrated loglikelihood, df = degrees of freedom.

the total twin sample). Furthermore, the mean age at last seen differs between the two groups (138.20 months for siblings and 151.48 months for twins). As the proportion of censoring in the sample increases so does the bias introduced by failing to account for the censoring in the statistical model. The difference between the censored siblings and censored twins with regard to mean age at last seen is also likely to increase the bias. After the removal of the siblings with an age at last seen less than 12 years old, the percentage of censored siblings dropped to 6.28%, with the mean age at last seen within the censored siblings increasing to 169.36 months. There remains a large difference (17.88 months) between the mean age at last seen of the siblings and twins, and this difference is due to the different ascertainment of the two groups. However, after the removal of siblings less than 12 years old, the total proportion of censoring is much smaller. Hence, the amount of bias introduced to the analysis when using a non-survival analysis method is expected to be significantly reduced by this treatment.

Further analyses were carried out to investigate the different ascertainment of siblings and twins, and its effects on the variance components estimates. The initial analysis, described above, selected only those twins with an age at interview of 12 years or more. A second analysis was carried out as described previously, but all siblings were removed from the sample prior to the estimation of variance components. A third analysis, which included all siblings regardless of age last seen, was carried out using the same methods. A summary of the results from all three analyses is given in Table 4.4.1.

When using COXME the siblings have little influence on the estimation of the variance components and the best fitting model, regardless of the inclusion criteria placed on the sibling sample, is an ACE model. This indicates that the COXME analysis is robust to the different censoring properties of siblings and

Table 4.4.1: Results from the additional analyses carried out to investigate the influence of the siblings on the variance components estimates from COXME and MX

	All Siblings		Siblings 12yr+		No Siblings	
	MX	COXME	MX	COXME	MX	COXME
Saturated model	ADE	ACE	ACE	ACE	ACE	ACE
A^2	0.53	0.56	0.76	0.57	0.50	0.54
C^2	-	0.24	0.06*	0.23	0.31	0.29
D^2	0.31	-	-	-	-	-
E^2	0.16	0.20	0.18	0.20	0.19	0.17

All Siblings includes every sibling in the sample, regardless of the age at last interview. Siblings 12yr+ summarises the results reported previously, where only those siblings with an age at last interview of 12 years or greater are included in the analysis. No siblings gives the variance components estimated by MX and COXME when only MZ and DZ pairs are included in the analysis. *Variance component is not significantly different from zero.

twins. However, the non-survival analysis method, carried out using MX, is not robust to the inclusion criteria placed on the sibling sample. If all siblings are included in the analysis the best fitting MX model is an ADE model; if all siblings are excluded from the analysis the best fitting MX model is an ACE model. The inclusion of the siblings clearly has a large effect on the variance component estimates when using a non-survival method. The only difference between the survival and non-survival methods is that the survival method correctly models the censored nature of the data in the statistical model, whereas the non-survival analysis does not. If all siblings are removed from the sample, there is agreement between the MX and COXME analyses, both with regard to the best fitting model (ACE) and the proportions of variance explained by these components. This suggests that the non-survival analysis is sensitive to the inclusion criteria of the siblings because of the different ascertainment and censoring seen in the siblings in comparison to the twins.

It was previously hypothesized that the bias introduced to the non-survival analysis estimation of variance components could be reduced by the selection of siblings with an age at last seen of 12 years or more. The results shown in Table 4.3.2 prove this hypothesis to be correct. Given that an ACE model is consistently the best fitting model when using a survival analysis method to estimate the variance components underlying variation in age at menarche, it is assumed that this is the best model to describe the variance in the current data. If the siblings with an age at interview of less than 12 years old are removed from the sibling sample, the non-survival analysis variance component estimates become closer to those estimated using the survival analysis method. The siblings are still introducing a small bias in the non-survival analysis, and this is only completely removed when the sibling sample is removed altogether.

The correlations (r^2) presented in Table 4.4.1 show the concordance in reported age at menarche to be greater in DZ pairs than in twin-sib pairs. This suggests a twin specific environmental effect on variance in reported age at menarche. However, because the correlations were estimated using MX, the censored nature of the data was not taken into consideration. Therefore, the different correlations reported for the DZ pairs and nontwin pairs could be due to the different censoring properties of the two groups. To investigate this further, a twin effect matrix was fitted using COXME (which accounts for the censoring present in the sample). If a significant difference exists between the correlations of the DZ pairs and non-twin pairs, the ACT model (additive genetic, common environment and twin specific variance components) will give a significantly better fit than the model where the twin specific environment effect has been dropped. The results of this investigation suggested that the twin specific environmental effect could be dropped from the model without significantly reducing the fit of the model ($P = 0.73$). It is therefore concluded that the difference in correlation between the DZ pairs and non-twin pairs shown in Table 3 are an artifact of the censoring properties of these two groups. The assumption that the means, variances and co-variance can be equated across zygosity groups appears to be correct.

Towne *et al.* (2005) review previous twin studies of age at menarche data and conclude that the heritability of age at menarche is approximately 0.50. This study provides evidence to support the findings of Towne *et al.* (2005). The components underlying variation in age at menarche are much more unclear. Of the 7 biometric studies discussed in Section 4.1, three different models (ADE, ACE, and AE) are given to describe variation in age at menarche. The present study suggests that the variation in age at menarche is influenced chiefly by additive genetic effects, with approximately 25% of the variance due to common environmental effects.

In summary, a biometric genetic analysis has been carried out on a sample of adolescent twins and siblings. The heritability of age at menarche was estimated to be 0.57 using a mixed effects Cox model. The analysis was carried out using a correlated frailty model to account for the 16.65% of individuals with a censored age at menarche. The analysis was also carried out without statistically accounting for the censored observations in the model, and when all siblings were removed from the analysis a heritability of 0.50 was reported. The best fitting model under both methods of analysis separated the variance in age at menarche into additive genetic (A), common environmental (C) and non-common environmental (E) effects.

Chapter 5

A genome-wide linkage scan for loci influencing variation in age at menarche

5.1 Introduction

As discussed in the previous chapter, age at menarche is a complex trait which is influenced by both genetic and environmental factors. The onset of menses is an important event both biologically and socially and the age at onset has been identified as a risk factor for several traits. Early age at menarche is a significant risk factor for several conditions, including depression (Kaltiala-Heino *et al.*, 2003), eating disorders (Kaltiala-Heino *et al.*, 2001), breast cancer (Velie *et al.*, 2006) and endometriosis (Missmer *et al.*, 2004). Late age at menarche has been associated with osteoporotic fractures (Naves *et al.*, 2005) and a reduced risk of coronary heart disease (Cooper *et al.*, 1999). The identification of QTL contributing to variation in age at menarche could potentially lead to a better understanding of factors underlying this wide range of phenotypes.

A secular decrease in age at menarche has occurred in developed countries over the past century (Ong *et al.*, 2006). This decrease is believed to be associated with the improved nutritional status, greater amounts of adipose tissue and the improved general health of adolescent females, though substances with

oestrogen-like actions that are present in nutrients have also been reported to play a role (deMuinich Keizer-Schrama and Mul, 2001). However, data from the Fels Longitudinal Study suggests that the population shifts in age at menarche and BMI are largely independent (Demerath *et al.*, 2004). This lack of association between BMI and age at menarche could be because BMI is an imprecise indicator of adiposity, particularly in children. It has been reported that the timing of menarche is more closely linked with body composition (i.e. body fat, fat percentage and lean tissue mass) during periods of growth and development. The Fels Longitudinal Study did find a correlation between adult BMI and reported age at menarche (Demerath *et al.*, 2004). There is evidence of a significant socioeconomic effect on age at menarche with means of 13.2 and 14.6 reported in privileged versus underprivileged black South African populations (Hughes and Kumanan, 2006). It is possible that the socioeconomic effect is due to an underlying nutritional effect on age at menarche as the two variables are likely to be highly correlated. A large multi-national study carried out by the World Health Organisation showed the median age at menarche to be 14 years though the analysis included many studies from developing countries. A mean age at menarche of 13 years was reported for the Australian sample (Morabia *et al.*, 1998).

The biological mechanisms through which puberty is mediated are only partially understood. Neurological control of the reproductive axis is driven by the hypothalamus, which has excitatory effects on the pituitary gland, leading to the release of follicle stimulating hormone (FSH) and luteinizing hormone (LH), which in turn regulate ovarian activity (Gottsch *et al.*, 2006). The initiation of menstrual cycles results from a series of developmental and neuroendocrine events that lead to full activation of the hypothalamic GnRH pulse generator and initiation of cyclic ovarian function (Tena-Sempere, 2006). Recent research has

focussed on the fundamental role of the Kisspeptin pathway in neuroendocrine regulation of the reproductive axis and initiation of puberty (Figure 5.1.1). Kisspeptin neurons in the anteroventral periventricular (AVPV) and arcuate (ARC) nuclei of the hypothalamus stimulate GPR54 receptors on GnRH neurons (Messager *et al.*, 2005).

A range of factors is involved in the control of puberty onset via the GnRH pulse generator, including leptin. Leptin is thought to provide information regarding the nutritional status of an individual to the hypothalamus and pituitary gland. Leptin, which is released from adipose tissue, has a stimulatory effect on both the ARC nucleus and the pituitary gland, leading to an increase in KISS neuron activation and circulating levels of LH and FSH, respectively (Yu *et al.*, 1997; Caprio *et al.*, 2001). Leptin could be underlying the correlation between body mass index and age at menarche.

To date, only two genome-wide linkage scans for genes underlying variation in age at menarche have been performed (Guo *et al.*, 2006a; Rothenbuhler *et al.*, 2006). Using 98 sister pairs Rothenbuhler *et al.* (2006) found three suggestive linkage peaks on chromosomes 16q21, 16q12 and 8p12 for weight-adjusted age at menarche. Three further suggestive QTL were identified on chromosomes 22q13, 22q11 and 11q23 following a linkage analysis for age at menarche carried out on a sample of 1,946 sister pairs (Guo *et al.*, 2006a). No replicated linkages have been reported for age at menarche and this is most likely due to a lack of powerful study designs; a study with a sample size of only 98 randomly selected sister-pairs will not have enough power to map loci underlying a complex trait such as age at menarche.

Several candidate genes have been associated with age at menarche, including the

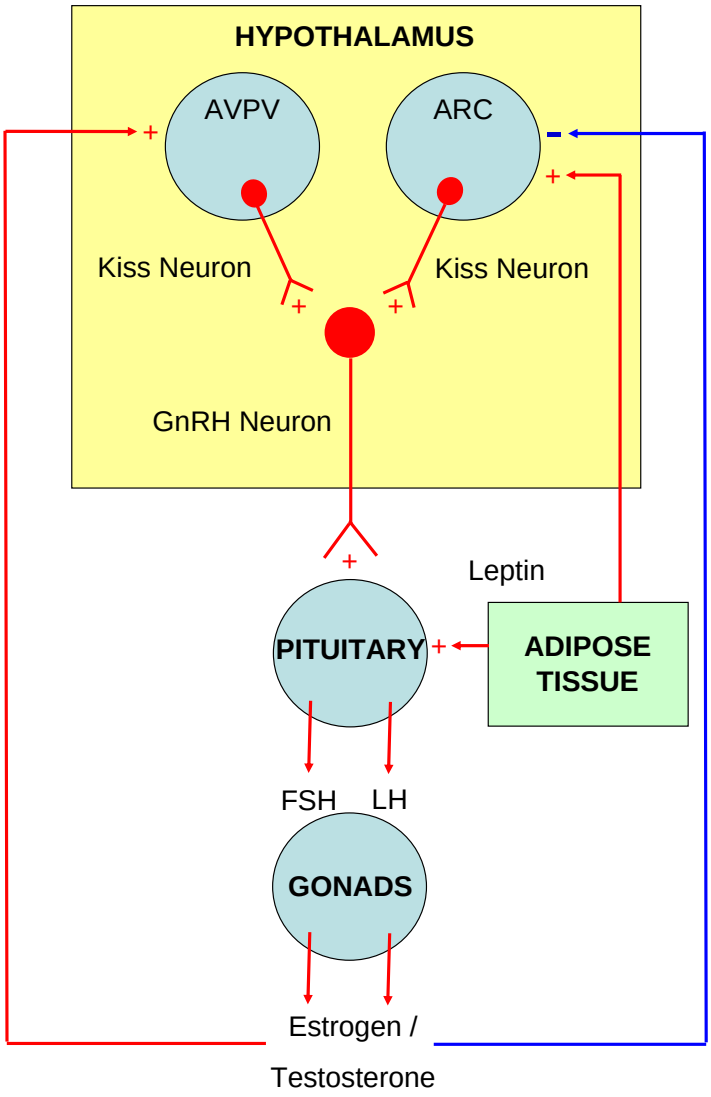

Figure 5.1.1: The neuroendocrine reproductive axis. Kisspeptin neurons in the anteroventral periventricular (AVPV) and arcuate (ARC) nuclei of the hypothalamus stimulate GPR54 receptors which stimulate gonadotropin-releasing hormone (GnRH) to be released from GnRH neurons. GnRH acts on the pituitary gland and initiates the release of follicle-stimulating hormone (FSH) and luteinizing hormone (LH). These hormones stimulate the gonads, which leads to an increase in the production of estrogen and testosterone, initiating puberty. The system is controlled via positive and negative feedback loops which act on the hypothalamus. Adipose tissue releases leptin which has a stimulatory effect on the ARC and pituitary gland and helps drive puberty.

estrogen receptor a (ESR1) gene (Stavrou *et al.*, 2002), the estrogen-biosynthesis gene CYP19 (Guo *et al.*, 2006b), and the sex hormone-binding globulin (SHBG) gene (Xita *et al.*, 2005). A recent review of CYP17 gene polymorphisms reported that 7 out of 11 studies suggested a modest association between CYP17 genotype and age at menarche (Sharp *et al.*, 2004).

In the present study, age at menarche data from four large-scale Australian study cohorts are combined to create the largest phenotypic sample of age at menarche to date. The aim is to conduct a highly-powered linkage analysis for loci underlying normal variation in age at menarche.

5.2 Methods

5.2.1 Phenotypic Sample

Age at menarche data were drawn from a cohort of adolescent twin families, from two cohorts of adult twin families and from a cohort of affected sister-pair families recruited for a linkage study of endometriosis.

Adolescent twin families: The adolescent twins and their families are decribed at length in the previous chapter. Of the 1,351 adolescent individuals in the sample, 226 (16.73%) had a censored age at menarche (i.e. the true age at menarche was unknown as the individual had not started puberty at the time of last interview). If censored observations are omitted from a sample then a bias is introduced when calculating means and variances and estimating heritabilities. In an attempt to reduce this bias the age at last seen was taken as the age at menarche.

Adult twin families: From 1980 to 1982, age at menarche information was collected from both members of 1,888 MZ and DZ female twin pairs as part of a health survey mailed to twins on the Australian Twin Registry. The cohort comprised of twins born between 1913 and 1964. Females were asked the question 'How old were you when you had your FIRST menstrual period?' (years and months) via questionnaire. A biometric analysis of age at menarche has previously been carried out on the sample (Treloar and Martin, 1990). From November 1989 onwards the first degree relatives of this twin cohort were also surveyed and asked the same age at menarche question. Age at menarche data from this cohort of female relatives are included in the present analysis. A second cohort of adult twins was recruited in 1989 through the Australian Twin Registry. The twins were born between 1964 and 1971, and hence were aged 18 to 24 years when they were initially surveyed by mailed questionnaire. Immediate family members were also asked to participate. As part of the questionnaire, participants were asked the question 'How old were you when you had your FIRST menstrual period?' (years and months). Non-responders from the second cohort were followed up, via interview, between 1996 and 2000 as part of a study into alcoholism and psychiatric morbidity (Heath *et al.*, 2001). Participants were again asked to provide an age at menarche. In total, the adult twin-families provided 6,150 families and 14,175 individuals with age at menarche information.

Endometriosis families: Age at menarche information was also obtained from women diagnosed with Endometriosis ascertained from the Australian component of the International Endogene Study (Treloar *et al.*, 2002). From 1995 to 2002, the Australian group recruited 931 families each with at least two affected members (mostly affected sister pairs) with surgically diagnosed endometriosis for a genome-wide linkage scan (Treloar *et al.*, 2005). Included

in the sample were case-parent trios and some cases without parents (Treloar *et al.*, 2002). All affected female family members were asked the question 'How old were you when your periods began' (whole years) via questionnaire. The total Australian sample provided 4,274 individuals with age at menarche information.

Given the small proportion of censored observations in the total phenotypic sample (1.20%) and the use of the age at last seen as an age at menarche in these cases, the bias in the present study as a result of censoring is predicted to be very small. For all study cohorts, where two or more age at menarche estimates were available for an individual, the estimate provided at the first data collection following menarche was used. It was assumed that the recall closest in time to menarche would be the most accurate. Age at menarche estimates less than 104 months (10 individuals) and greater than 208 months (64 individuals) were classified as outliers and removed from the analysis. These cut-offs approximately represented three standard deviations above and below the mean. The total phenotypic sample for the present linkage analysis comprises of 19,782 individuals with an age at menarche. Phenotypic bivariate outliers were identified and removed to ensure that linkage signals were not overly influenced by extremely discordant sib-pairs. A bivariate outlier is defined as a sib-pair with extreme differences in age at menarche. The distribution of the squared difference in age at menarche between pseudo-independent sib-pairs is given by

$$\frac{D^2}{2(1-r)\sigma^2} \sim \chi^2_1, \tag{5.2.1}$$

where r is the sib-correlation and σ is the standard deviation, of the trait (age at menarche). It follows from this that the mean of D^2 is given by

$$\mathrm{E}(D^2) = 2(1-r)\sigma^2, \tag{5.2.2}$$

and the variance of D^2 is given by

$$\sigma^2(D^2) = 8(1-r)^2\sigma^4. \tag{5.2.3}$$

Hence, the expected standard deviation of the squared difference is

$$\sigma(D^2) = 2\sqrt{2} \times (1-r)\sigma^2. \tag{5.2.4}$$

Pseudo-independent sib-pairs with a D^2 greater than $4\sigma(D^2) + \mathrm{E}(D^2)$ were identified as bivariate outliers. The sib-pair correlation for age at menarche calculated by SIB-PAIR (Duffy, 2002) was 0.28. The standard deviation of age at menarche from the full phenotypic sample was 17.20 months (1.43 years). Using equation 5.2.4 the expected standard deviation of the squared difference between pseudo-independent sib-pairs was 602.47 months^2. Using equation 5.2.2 the mean of D^2 equals 426.01 months^2. Therefore, any pseudo-independent sib-pair with a squared difference in age at menarche of more than 2835.88 months^2 was identified as a bivariate outlier. This is equivalent to an actual difference of $\sqrt{2836}/12 = 4.44$ years between the age at menarche of pseudo-independent sib-pairs. For each pseudo-independent sib-pair identified as a bivariate outlier the age at menarche which differed most from the sample mean was removed from further analysis.

5.2.2 Genotypic Sample

Adolescent twin families: Three genome-wide scans using microsatellite markers were carried out on extracted DNA from blood samples collected from the adolescent melanoma and cognitive ability families. Details of the genotyping and genetic data cleaning procedures have been described previously (Zhu *et al.*, 2004). Genetic data from all three genome screens are used in the

present study. The adolescent samples provided 628 individuals with both age at menarche and genotypic data to this present analysis.

Adult twin families: Four microsatellite genome-wide scans have been completed on DNA extracted from blood or buccal samples collected from individuals within the two adult cohorts. The genotyping and genetic data cleaning procedures for all four scans have been described, at length, previously (Cornes *et al.*, 2005). The adult cohorts provided 3,544 individuals with both phenotypic and genotypic data to the present study.

Endometriosis families: Details of the genotyping and genetic data cleaning procedures for the endometriosis cohort have been detailed elsewhere (Treloar *et al.*, 2005). A single 400 marker genome-wide screen was carried out on the Australian families. In a time and money saving exercise, the final batch of 79 families was genotyped using only 113 markers across chromosomes 9, 10, 11, 19, 20, 21, 22 and X. The endometriosis cohort provides 2,160 individuals with both phenotypic and genotypic data to the present study.

In total, the present study comprises 6,332 individuals and 2,685 pseudo-independent sib-pairs with age at menarche and genotype data. This is the largest sample of pseudo-independent sib-pairs used to date in a genome-wide scan for loci influencing age at menarche. Of the 2,685 pseudo-independent sib-pairs in the linkage sample 42 were identified as bivariate outliers (i.e. had a difference in age at menarche of more than 4.44 years). After the removal of bivariate outliers the linkage sample consisted of 6,290 individuals and 2,631 pseudo-independent sib-pairs.

In addition to the initial checks of genetic data which were carried out during

the original studies, further checks were applied following the amalgamation of cohorts. The data were re-checked for Mendelian errors using SIB-PAIR (Duffy, 2002). Unlikely genotypes were identified and wiped using MERLIN (Abecasis *et al.*, 2002).

This study uses an updated version of a previously described (Nyholt *et al.*, 2005) genetic map which takes physical map positions from NCBI build 35.1 and genetic map positions from published deCODE and Marshfield data (Kong *et al.*, 2002; Myers *et al.*, 2005; Duffy, 2006). Where the same marker was typed in several genome-screens, the multiple marker versions were offset by 0.002cM to prevent them sharing the same map position and to negate the need to re-bin alleles across all studies.

5.2.3 Statistical analysis

The mean and variance of age at menarche was calculated both within and across all studies. Genome-wide variance-component (VC) linkage analysis was carried out using MERLIN (Abecasis *et al.*, 2002). VC linkage analysis is a model-free analysis method, i.e. penetrance parameters and disease gene frequencies are unspecified. Identity-by-descent (IBD) coefficients were calculated at 1cM intervals. The IBD coefficient gives the number of alleles two individuals share, at a given locus, which are derived from the same ancestor (0, 1 or 2 alleles). Variance-component models are based on the correlation between the genetic similarity of relatives at a given locus (estimated IBD coefficients) and the relatives' similarity with respect to the phenotype (age at menarche). Phenotypic data were assumed to follow a multivariate-normal distribution. MERLIN allows the inclusion of phenotypic information from both members of a monozygotic twin pair provided the individuals are input as a monzygotic twin pair. In the present

linkage scan, age at menarche data from monozygotic twins was included, which increases the power to detect linkage (Evans and Medland, 2003). In accordance with previously proposed significance levels (Lander and Kruglyak, 1995) for model-free genome-wide linkage analysis, a genome-wide significance threshold of LOD = 3.3 was adopted. Suggestive evidence of linkage was reported when a LOD score greater than or equal to 1.9 was observed. The linkage analysis was carried out both before and after the removal of bivariate outliers. It is worth highlighting that the full pedigree structure is fitted in the variance component model and that the dependency between pseudo-independent sib-pairs is therefore fully accounted for. The term 'pseudo-independent pair' is only used to give an approximate estimate of power and to identify within-family bivariate outliers.

5.3 Results

The mean and variance of age at menarche within and across studies is shown in Table 5.3.1. After combining phenotypic data from all four studies the mean and standard deviation of age at menarche was 13 years and 1.42 years, respectively. Figure 5.3.1 shows the distribution of age at menarche estimates following the removal of outliers. The means and variances reported here are similar to those reported in previous studies (Towne *et al.*, 2005). Variance-component analysis showed the heritability of age at menarche within our study to be 64% (where all familial resemblance is assumed to be additive genetic), which is similar to levels of heritability reported in other studies (Towne *et al.*, 2005; Treloar and Martin, 1990).

Results of the genome-wide variance-component linkage analysis are shown in Figure 5.3.2 along with the locations of candidate genes. Two primary candidate genes, INS and GPR54 are included even though they are located at the extreme

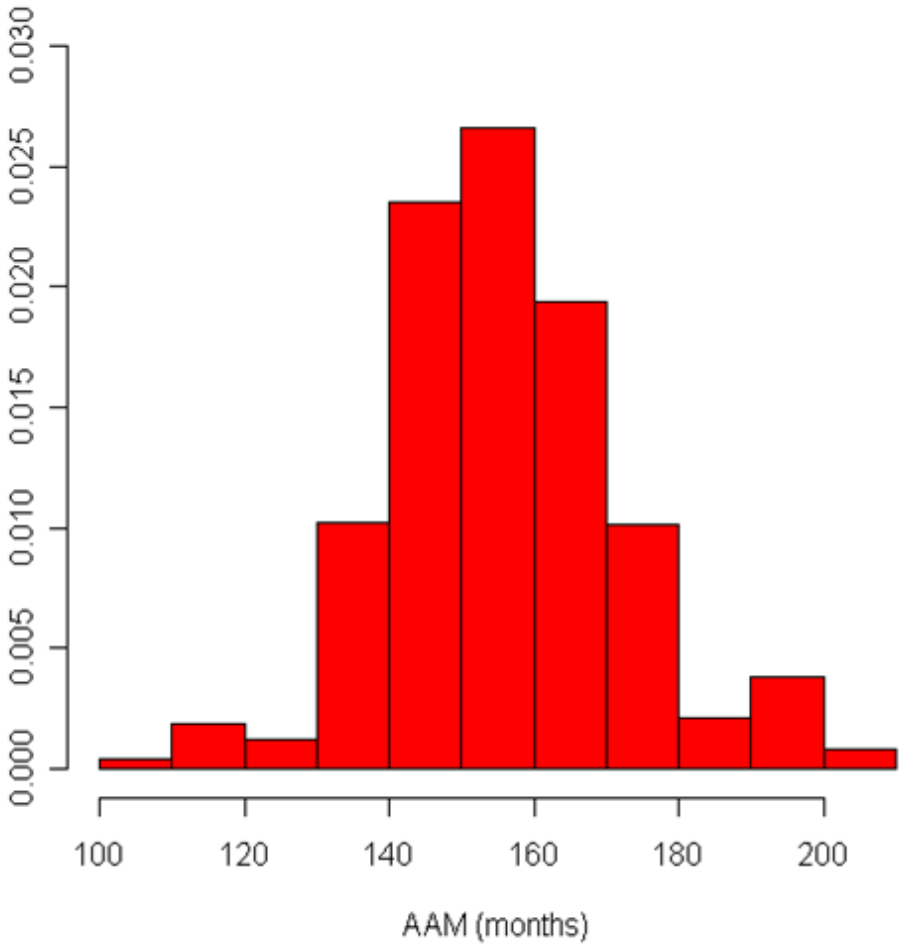

Figure 5.3.1: Histogram of age at menarche data from 19,782 Australian females

Table 5.3.1: Means and standard deviations of age at menarche from four different Australian cohorts

Study	Individuals	Mean (months)	Standard deviation (months)
Adolescents	1345	154	14
Adults 1 (1980 cohort)	9099	157	17
Adults 2 (1989 cohort)	5076	158	17
Endometriosis	4262	154	18
Combined	19782	157	17
Linkage subset	6332	156	17

ends of chromosomes 11 and 19, respectively, and are outside the resolution of the current genome-wide scan. If either gene had a major effect on the trait one would still expect to see a LOD peak at the start of our marker region on chromosome 11 or 19, and this is not the case. Whilst this does not rule out INS or GPR54 as genes with effects on age at menarche, it is unlikely they have a large effect in the current population. Regions of suggestive linkage (LOD = 1.9) were found on chromosomes 3 and 11, with LOD scores of 2.31 and 2.16, respectively. The linkage peak on chromosome 3 was found at a map position of 10.2cM, located on the p-arm of the chromosome between markers D3S1297 and D3S1620. The linkage peak on chromosome 11p was located at a map position of 56.9cM, between markers D11S4200 and D11S1900. Both of these suggestive linkage regions are novel loci for genes influencing age at menarche. A summary of the five most significant loci identified by genome-wide linkage analysis is provided in Table 5.3.2.

The LOD score profile across the chromosomes changed significantly after the removal of bivariate outliers. The LOD score peaks on chromosomes 3 and 11

Table 5.3.2: Summary of the five most significant loci from multipoint variance-component linkage analysis for age at menarche

Chromosomal Band	Location	LOD score	Flanking markers	
			Upstream	Downstream
3p26	10	2.31	D3S1297	D3S1620
11p13	57	2.16	D11S4200	D11S1900
12q24	133	1.77	D12S79	D12S86
9q21	84	1.18	D9S167	D9S257
12q21	96	1.13	D12S106	D12S379

decreased from 2.31 and 2.16 to 0.49 and 1.11, respectively, and are no longer of suggestive significance. This indicates that the removed pseudo-independent sib-pairs provided a large proportion of the information for linkage at these loci. Perhaps more encouraging is the LOD score peak seen on chromosome 12 which did not change after the removal of the bivariate outliers. This suggests that the removed individuals had little influence on the LOD score at this test locus. In addition, a larger peak with a LOD score of 1.88 is seen on chromosome 12 at a position of 154cM, which suggests the bivariate outliers had a negative effect on the LOD score profile in this genomic region. Other chromosomes with increased lod score peaks include chromsomes 8, 17 and 22, which after the removal of bivariate outliers show LOD score peaks of 1.27, 1.75 and 1.52, respectively.

5.4 Discussion

In this study, a genome-wide scan for loci influencing age at menarche was conducted using a sample of 6,332 Australian females. Multipoint variance-component linkage analyses revealed suggestive linkage on chromosome 3p26 (LOD = 2.31) and 11p13 (LOD = 2.16).

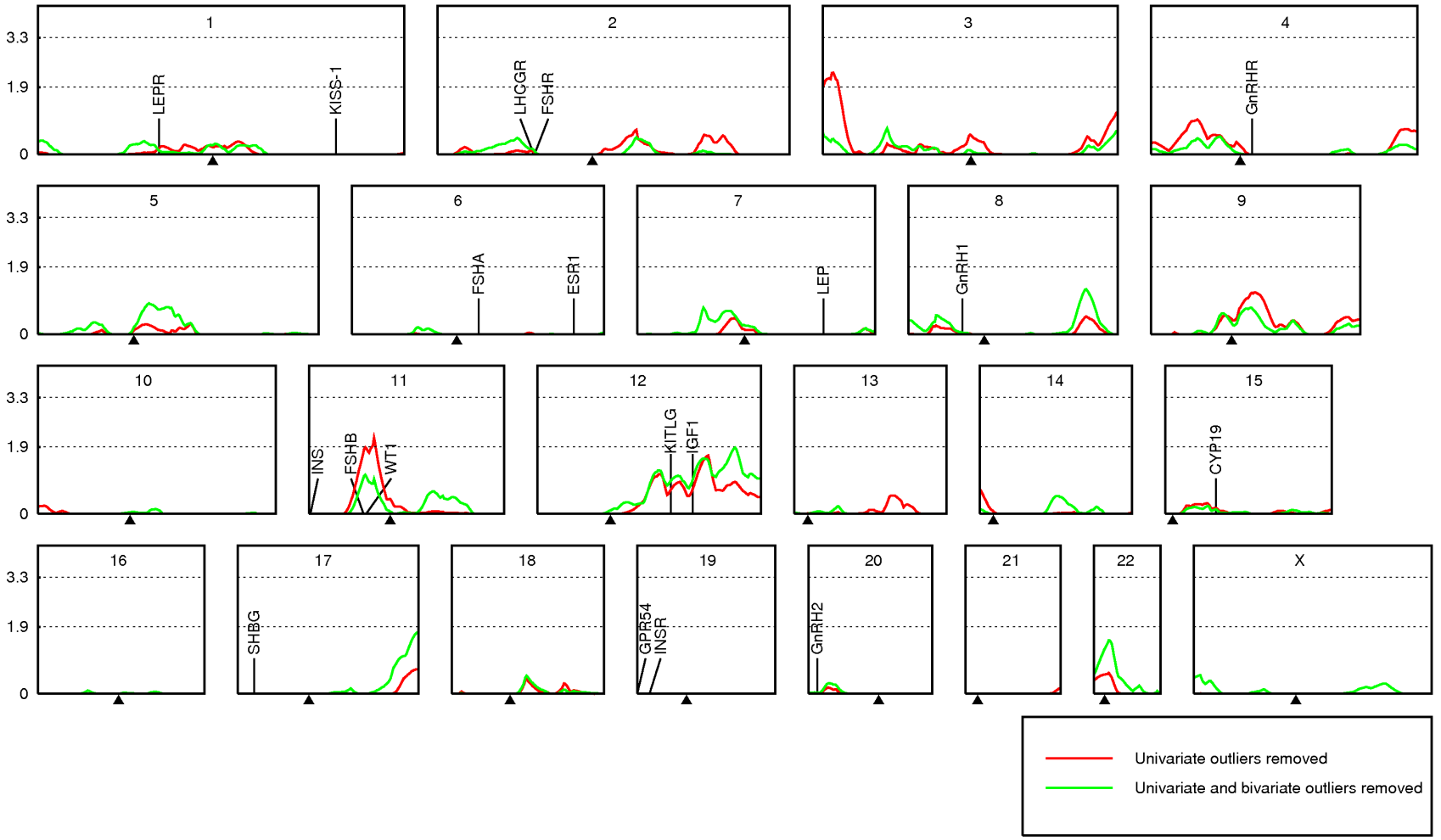

Figure 5.3.2: Genome-wide variance component LOD scores by chromosome. The centromere position is given by ▲

A search for candidate genes was carried out within the confidence regions of the linkage peaks on chromosomes 3 and 11 (see Figure 5.3.2). The confidence region of a QTL is defined as the region of the chromosome with a LOD score = peakLOD - 1. No candidate genes were identified within the QTL confidence region on chromosome 3. The confidence region on 11p harbors two excellent candidate genes for age at menarche. Wilms Tumour 1 (WT1 [OMIM: 607102]) is a 48,000 base-pair gene located on chromosome 11p13. WT1 is required for the normal formation of the genitourinary system and mesothelial tissues. WT1 is expressed in immature follicles and has been associated with follicular development. In WT1 knockout mice gonadal development is inhibited, which indicates that WT1 is necessary for the differentiation of the gonadal ridge (Kreidberg *et al.*, 1993). Individuals with Wilm's Tumour, or the related Frasier syndrome and Denys-Drash syndrome, are often infertile. The role of WT1 in follicular development and the syndromes associated with WT1 mutations suggest that WT1 is a prime candidate for a gene influencing age at menarche.

The second candidate gene within the chromosome 11p linkage region is Follicle stimulating hormone - beta polypeptide (FSHB [OMIN: 136530]), a 2,278 base pair gene located on chromosome 11p13. FSH drives ovarian folliculogenesis to the antral follicle stage, and individuals with no FSH-receptor activity have been shown to have hypergonadotrophic primary amenorrhea and hyperplastic ovaries (Huhtaniemi, 2002). FSH plays a critical role in the neuroendocrine axis and induces estrogen to be released from the ovaries (see Figure 5.1.1). Concentrations of FSH are known to increase during normal pubertal development. Several studies have reported mutations within FSH as the causal variants for amenorrhea. A two base pair frameshift deletion was found in the coding sequence of the FSH gene from two women, one Italian

and one Israeli, with primary amenorrhea and infertility associated with an isolated deficiency of pituitary FSH (Matthews *et al.*, 1993; Matthews and Chatterjee, 1997). A 16 year old female with amenorrhea who had undergone pubarche (pubic hair development) but not thelarche (breast development) was shown to be heterozygous for two FSH mutations. Each of the mutations prevents the normal combination of the alpha and beta subunits to form intact FSH. The authors conclude that the findings indicate that FSH is required for follicular development and ovarian androgen and estrogen synthesis in females (Layman *et al.*, 1997). These findings together suggest that the FSH gene is an excellent candidate for a gene with influence on age at menarche.

After the removal of bivariate outliers the LOD score profile changed substantially. This is not unexpected as extremely discordant sib-pairs are very informative for linkage. For a chromosomal region where an extremely discordant sib-pair share zero or two alleles identical-by-descent that sib-pair will have a large influence on the LOD score. Where they share zero genes identical-by-descent the sib-pair will have the effect of increasing the LOD score and where they share two genes identical-by-descent they will decrease the LOD score. In the perfect scenario a LOD score peak is influenced by multiple sib-pairs and multiple families, but this is seldom the case as the extremely concordant and discordant sib-pairs provide the majority of the power to detect linkage. For this reason, it is difficult to know how to correctly handle bivariate outliers in linkage analysis, hence why the analysis is carried out both with them included and excluded. It seems reasonable to suggest that one may have more confidence in a linkage peak that is still present in the absence of bivariate outliers (chromosome 12), than those that are seen only in their presence (chromosomes 3 and 11). Furthermore, the aim of this study was to identify loci underlying the normal variation in age at menarche, and therefore it would seem wise to pay particular attention to the

linkage results following the removal of bivariate outliers.

The factors leading to bivariate outliers remain unclear, but one possible explanation is that the age at menarche for one individual within the pseudo-independent sib-pair is incorrect. This could, in part, be due to recall or measurement error on behalf of one or both members of the twin pair. The removal of bivariate outliers could therefore be removing the most extreme incidents of recall bias. Another, and perhaps more interesting, explanation is that the individual within the pair with the most extreme age at menarche is segregating a polymorphism of major effect upon the trait. The timing and control of puberty is believed to be driven by the hypothalamus and so primary candidate genes are thought to act at this level of the pubertal axis. The linkage peak around FSHB decreases after the removal of bivariate outliers, and it is known that FSHB acts at the level of the ovary, not the hypothalamus. Genes with effects on the ovaries are not expected to have control over the normal regulation of pubertal timing, but they could have an effect on the ability of a female to ovulate. This is supported by the studies that reported amenorrhea in women with mutations in FSHB (Matthews *et al.*, 1993; Matthews and Chatterjee, 1997; Layman *et al.*, 1997). An explanation of the decrease seen on chromosome 11 is that the individuals removed as bivariate outliers are all segregating polymorphisms within the FSHB gene that have a major effect on a womens ability to ovulate, and therefore have a major effect on age at menarche. The polymorphisms are not involved in the pathways that control the timing of menarche per se.

Despite the large number of sib-pairs the present study reveals no genome-wide significant linkage peaks, suggesting that the genetic architecture of age at menarche is complex. If a single QTL explained the variation in age at menarche

then we would expect to detect a gene of this effect size in a study of this magnitude. The fact that we have not detected a significant LOD score suggests that multiple loci are involved in the heritability of age at menarche. This finding alone is rather unremarkable; however we would still have 93% or 67% power to detect a QTL (at $\alpha = 0.0001$, genome-wide significance) that explains 25% or 20% of the trait variation, respectively (Sham *et al.*, 2000; Purcell *et al.*, 2003). This suggests that age at menarche variation is underpinned by multiple QTL of small effects and is similar to the genetic architecture of other complex traits (Reich and Lander, 2001). Large numbers of sib pairs are needed to identify QTL with small effects using linkage analysis and this would explain why no statistically significant QTL were detected in the present linkage scan. If a QTL existed that explained 10% of the trait variation, for a trait with a heritability of 0.65 such as age at menarche, then 13,000 pseudo-independent sib-pairs would be required to detect the QTL with 80% power. This number of sib-pairs is outside the scope of most genome-wide linkage studies. The recent availability of genome-wide association study designs should allow the detection of these small effect QTL (Sham *et al.*, 2000). Rare alleles with large effects on age at menarche could still be segregating within the population; however, due to their rarity, it is unlikely that they will be segregating within our population sample. Furthermore, due to their rarity they would explain little of the population variation in the trait.

It has been suggested that including individuals without genome-wide linkage information in a genome-wide scan for linkage can bias the result towards the null hypothesis (Schork and Greenwood, 2004), though this claim has been hotly contested (Abecasis *et al.*, 2004; Cordell, 2004; Mukhopadhyay *et al.*, 2004; Sieberts *et al.*, 2004; Visscher and Wray, 2004). To look at this empirically, the genome-wide linkage analysis was repeated using only those

individuals with genome-wide linkage information. We define an individual as having genome-wide linkage information if they share 300 or more typed markers in common with a sibling. Of the 6,332 individuals and 2,685 pseudo-independent sib-pairs with age at menarche and genotype data, 5,700 individuals and 2,482 pseudo-independent sib-pairs have genome-wide marker data available. There were no significant differences between the linkage results of the two analyses. The maximum difference between the LOD scores of the two linkage scans was 0.21, with the higher LOD score resulting from the analysis which included all individuals, regardless of marker information. This suggests that including individuals with 'missing' inheritance information (i.e. IBD = 0.5) does not bias the test statistic towards the null hypothesis of no linkage. Furthermore, by excluding those individuals with less than 300 typed markers shared in common with a sibling, some linkage information is being lost.

In summary, the largest genome-wide scan for loci influencing variation in age at menarche has been carried out. Two novel loci of suggestive significance have been identified on chromosomes 3p and 11p. The linkage region on chromosome 11 contains WT1 and FSHB, two candidate genes for age at menarche. Chromosomes 12 and 17 contain loci which just failed to reach suggestive significance, and these loci were identified in the absence of bivariate outliers. Given the size of the sample it would appear there are many genes of small effect contributing to variation in age at menarche.

Chapter 6

Discussion

Elucidating the genetic basis of phenotypic variation is the foremost goal in the field of genetics. The much heralded completion of the human genome project in the year 2001 signalled a new era for 'gene mappers' (The International Human Genome Sequencing Consortium, 2001). Subsequently, many complex traits have been the subject of gene mapping experiments, particularly those contributing significantly to the global burden of disease. However, perhaps due to the complexity involved in the analysis, age at onset data have received much less attention than standard quantitative traits. The major aim of this thesis was to identify loci underlying quantitative traits through the use of age at onset information.

6.1 Thesis summary

Age dependent penetrances were included in a marker-specific parametric linkage analysis of four families with major depression and comorbid unexplained swelling (Chapter 2). Parametric linkage analysis was used because, due to the disease segregation patterns, it was assumed that a Mendelian locus of major effect was segregating within the families (albeit with incomplete penetrance). The use of parametric linkage analysis allowed the age-specific penetrances to

be modelled in unaffected individuals. This approach is perhaps the simplest way in which the 'age at onset' distribution of a trait can be included in a genome-wide scan for linkage. A locus of suggestive linkage was identified on chromosome 8q with a LOD score of 2.02. The locus was not identified in a follow-up non-parametric variance component genome-wide scan, though a second locus of suggestive linkage was mapped to chromosome 7 (LOD = 2.10). Due to the novel comorbidity phenotype, the sample size was fixed, and after maximization of marker information a previously suggestive linkage on chromosome 14 disappeared. Future efforts should be directed at ascertaining more families with this novel phenotype in an attempt to increase the power to detect linkage. A more thorough examination of the novel comorbid phenotype would also aid the hunt for underlying susceptibility loci by increasing the accuracy of parametric linkage models.

Standard quantitative trait loci (QTL) mapping techniques commonly assume that the trait is both fully observed and normally distributed. When considering survival or age-at-onset traits these assumptions are often incorrect. Methods have been previously developed to map QTL for survival traits; however, they are both computationally intensive and not available in standard genome analysis software packages. A grouped linear regression method is proposed for the analysis of continuous survival data in line crosses (Chapter 3). Using simulation the method is compared to both the Cox and Weibull proportional hazards models and a standard linear regression method that ignores censoring. The grouped linear regression method is of equivalent power to both the Cox and Weibull proportional hazards methods and is significantly better than the standard linear regression method when censored observations are present. The method is also robust to the proportion of censored individuals and the underlying distribution of the trait. Based on linear regression methodology, the

grouped linear regression model is computationally simple and fast and can be readily implemented in freely available statistical software. The extension of the method to allow the modelling of survival traits in general outbred populations is an important next step. As previously suggested, a random-effects QTL model based upon multiple binary indicator variables would naturally fit into a linear mixed model framework. It is likely that this linear regression based methodology will produce even greater savings in computation time versus current methods for mapping survival QTL in outbred populations.

Prior to a gene mapping experiment, a biometric analysis should be carried out to determine the proportion of variance explained by genetic effects. For age at onset traits, this usually straightforward step is complicated by the need to use specialized methods to account for the non-normal distribution of the data and the censored observations. In Chapter 4, a standard variance component method was compared to a corresponding survival analysis method using a sample of 1,302 adolescent twins and their siblings for whom age at menarche information was available. The proportion of censoring in the sample was 16.73%. The best fitting model following analysis with the survival analysis method was an ACE model, where 57% and 23% of the phenotypic variance was explained by additive genetic and environmental effects, respectively. The best fitting model when using a standard variance decomposition method was an AE model, where 82% of the phenotypic variance was explained by additive genetic effects. The phenotypic data were investigated further and the lack of correspondence between the results of the two methods was found to be an artefact of the different ascertainment of reports of age at menarche between the siblings and twins. After the removal of the sibling sample, the correspondence of the two method increased, with both methods indicating an ACE model was the most likely. The standard and survival analysis methods estimated the proportion

variance explained by additive effects to be 0.50 and 0.54, and the proportion variance explained by environmental effects to be 0.31 and 0.29, respectively. These results suggest that great care must be taken when using non-survival analysis methods to decompose the variation in survival traits. Further work is needed to accurately define the conditions under which survival analysis methods must be used in place of more traditional schemes.

Only 628 adolescent twins and their siblings had both age at menarche information and marker data available. For a complex trait such as age at menarche, where variance is likely to be underpinned by many modest or small effect QTL, this sample size is too small. In an attempt to increase the power to detect linkage, the sample was broadened to include adults for whom age at menarche data was available. The total linkage sample comprised of 2,685 pseudo-independent sib-pairs with age at menarche and genotype data. This represents the largest sample of pseudo-independent sib-pairs ever ascertained for a genome-wide linkage scan for loci underlying variation in age at menarche. Two QTL of suggestive significance were identified on chromosomes 3 and 11, with LOD scores of 2.31 and 2.16 respectively. After the removal of bivariate outliers the evidence for linkage at both of these QTL decreased. However, a third QTL was identified on chromosome 12 (LOD = 1.88). All of these QTL are novel loci for age at menarche, though only two previous genome-wide scans for loci influencing age at menarche have been completed to date. The linkage region on chromosome 11 contains two candidate genes for age at menarche, WT1 and FSHβ. Given the sample size of the present study and that no significant QTL were identified, it seems likely that age at menarche is influenced by multiple genes of small to medium effect. Therefore, a very large number of individuals will be required to map QTL for age at menarche using a linkage analysis approach. This finding is in common with those of the majority of complex trait QTL mapping studies.

6.2 Future directions for gene identification

Five years after the completion of the human genome project and still only a handful of the polymorphisms underlying complex human traits and diseases have been unequivocally identified. As discussed in Chapter 1, linkage analysis has been a lucrative method for identifying loci underlying Mendelian traits, though much less has been achieved with complex traits. The lack of success is primarily because the majority of studies not only select heterogeneous samples but have insufficient power to even detect QTL of modest effect size (Altmuller *et al.*, 2001). The proportion of significant complex trait linkages that have been replicated in several independent studies is very small. Even after a linkage signal has been replicated across several independent studies, the underlying causal polymorphism is often still difficult to localize because many candidate genes can reside within the confidence region of the QTL due to the extent of linkage signals (typically tens of centimorgans for complex traits). Given the lack of success in identifying polymorphisms underlying complex trait variation using traditional linkage analysis methods, what does the future hold for gene mappers?

With the mapping of QTL underlying complex traits in mind, several techniques for improving the efficacy of traditional linkage analyses have been suggested. In addition, recent advances in technology have presented several new methods for identifying the underlying causes of complex trait variation.

6.2.1 Improving the design of genetic linkage studies

Inter-study variation in study design, phenotype definition, analysis methods and selection of genetic markers all reduce the chances of finding a replicated linkage for a complex trait. When these factors are coupled with interpopulation heterogeneity and differences in environmental exposures it is perhaps not

surprising that few complex trait linkage reports have been consistently replicated across populations.

Phenotype definition is an important issue in complex trait gene mapping. As discussed previously, locus heterogeneity can be reduced by ascertaining narrowly defined subsets of disease phenotypes (Chapter 2). The ascertainment of families with early or late onset, comorbities, multiple affected individuals or similar disease severities are all classic ways to reduce locus heterogeneity. However, the assumption that this method will actually increase the number of replicated linkages is open to debate. As study designs place more and more selection criteria on linkage samples, the chance of a similar sample being ascertained in an independent study is surely decreased.

When ascertaining probands for genetic studies, researchers typically depend on phenotype descriptions from the field of medicine. For example, the definition of major depression used in Chapter 2 is taken from the Diagnostic and Statistical Manual of Mental Disorders IV (American Psychiatric Association, 1994), a manual which serves as a guide to clinicians. While these descriptions are useful for identifying individuals with a particular ailment, they were never intended to identify individuals with genetically similar characteristics. A solution to this problem would be the creation of a set of phenotype specific guidelines for genetic epidemiologists to follow when ascertaining individuals for a given trait. Amongst other criteria, the guidelines should give clear definitions of early and late onset, disease thresholds (e.g. high versus low blood pressure), familial versus non-familial forms of the disease and any comorbid phenotypes that should be included. The guidelines would allow independent studies to ascertain phenotypically similar subsets of disease. This would not only increase the chances of finding replicated evidence for linkage, but also allow joint analysis

(meta-analysis) across studies in an attempt to increase power.

The issue of power is a significant problem in complex trait linkage analysis. Only recently have researchers become aware of the number of individuals that must be ascertained to identify QTL underlying complex trait loci. For complex diseases it seems likely that the effect size of any one polymorphism is likely to be small. The number of pseudo-independent sib-pairs required to map these loci of small effect using a linkage framework will run into the tens of thousands (Chapter 5). In an attempt to ascertain these large number of individuals, several large biobanks have been set up worldwide, and include Biobank UK (http://www.ukbiobank.ac.uk/), Generation Scotland (www.generationscotland.org), EPIC Europe (http://www.iarc.fr/epic/centers/iarc.html), GenomeEUtwin (http://www.genomeutwin.org), The Western Australian Genome Project (http://www.genepi.com.au/wagp) and the Decode Genetics Icelandic Biobank (http://www.decode.com). Biobanks involve the storage of biological and phenotypic data for large population-based samples. The sample sizes of such studies are typically in the order of hundreds of thousands, and the largest population Biobank, the Western Australian Genome Project, includes in excess of 2 million people. Biobanks which collect family information, such as the The Western Australian Genome Project, provide an excellent resource for linkage analysis of complex traits, particularly as the study protocol is consistent for the entire sample. Population biobanks are an excellent resource for many different phenotypes, and the data they contain can be put to good use in many different fields, including human genetics. However, for less common diseases, the size of the population samples needs to be very large to ensure the ascertainment of enough affected individuals. How to make best use of this data for the genetic analysis of complex traits is an ongoing debate, with many authors suggesting

alternatives to traditional linkage analysis. The number of individuals required to successfully map a QTL of small effect using linkage analysis rules out the method in all but the largest of studies, and therefore alternative methods must be adopted. One such method which has received much attention of late is genome-wide association mapping.

6.2.2 Genome-wide association studies

The traditional alternative to a linkage study is a hypothesis-driven candidate gene association study. The advantage of an association study approach is that a smaller number of individuals is required to detect loci of small effect (Risch, 2000). As shown in Table 6.2.1, for the detection of small effect (quantitative trait nucleotide (QTN) explains 1% of the phenotypic variance) linkage analysis using sib-pairs requires approximately 400 times the number of individuals needed for an association analysis. In Table 6.2.1, the number of individuals required to map a QTN with 80% power using association was calculated assuming r^2 of 1 between the marker SNP and QTN. This assumption is optimistic given the current number of SNPs being used in genome-wide association studies (typically about 500,000). An r^2 of 0.8 between the QTN and marker SNP seems more likely, and therefore a greater number of individuals (approximately 1/0.8 more) than that shown in Table 6.2.1 is currently needed to map a QTN using an association approach. Furthermore, if the LD structure is taken into account when selecting tagging SNPs, and the fewest number of SNPs are chosen to tag the entire genome (i.e. there is little or no redundancy with regard to SNP markers) then genotyping error or failure can have large impacts on the power to detect association in the given genomic region.

Typically, an association approach is the method of choice to detect common

variants of small effect, while linkage analysis has more power to detect rare loci of large effect segregating in large pedigrees. Therefore, the genetic architecture of the trait must be fully considered when deciding which method, linkage or association, is likely to be most powerful for a given study. In addition, the extent of linkage disequilibrium, the basis of association mapping, is much less than the confidence region surrounding a linkage peak. Therefore, the resolution of an association is much better than that of linkage (Cardon and Bell, 2001). Consequently, a greater number of markers are needed to cover a genetic region when carrying out an association study, and as a result studies have typically been limited to a few genes or a particular genomic region. Since the completion of the human genome project (The International Human Genome Sequencing Consortium, 2001), the mapping of 1.4 million single nucleotide polymorphisms (SNPs) (The International SNP Map Working Group, 2001), the ongoing progress of the International HapMap Project (The International HapMap Consortium, 2003; The International HapMap Consortium, 2005) and the advent of moderately low-cost, high-throughput genotyping technology, hypothesis-free genome-wide association (GWA) studies involving hundreds of thousands of SNPs genotyped on thousands of individuals is now a realistic aim. Currently, there are almost 12 million confirmed *Homo sapien* SNPs in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/, Build 126).

Showing incredible foresight, Risch and Merikangas (1996) proposed a genome-wide association design long before today's technology made it a realistic option for the analysis of complex phenotypes. Many research groups worldwide have plans to use the method to elucidate polymorphisms underlying variation in a plethora of human disease phenotypes (Thomas *et al.*, 2005). Results from initial genome-wide association studies are starting to appear in the literature, some with encouraging results (Ozaki *et al.*, 2002; Klein *et al.*, 2005; Duerr

Table 6.2.1: Power to detect QTL using either a linkage or association approach

Phenotypic variance explained by QTL	Linkage N (80% power)	Association N (80% power)	Ratio $\left(\frac{linkage}{association}\right)$
0.20	6,832	308	22
0.10	27,480	654	42
0.05	110,060	1,342	82
0.01	2,752,000	6,852	402

N (80% power) gives the number of individuals required to map a QTL with 80% power at a given QTL effect size (calculated using the online GENETIC POWER CALCULATOR (http://pngu.mgh.harvard.edu/ purcell/gpc/) (Sham *et al.*, 2000; Purcell *et al.*, 2003)). The linkage power analysis was carried out assuming a heritability of 50%, and a type I error rate of 0.0001. A sibship study design was assumed. The association power analysis was carried out assuming a type I error rate of 0.0000005 (the 95% significance threshold following a bonferroni correction when testing 1 million SNP markers), a singleton study design and that the marker and QTN are in perfect linkage disequilibrium ($r^2 = 1$).

et al., 2006; Smyth *et al.*, 2006).

The design and analysis of genome-wide association studies has received much attention (Cardon and Bell, 2001). Many options are currently available to researchers in terms of design, including family-based or case-control samples, map-based or gene-based SNPs, tagging SNPs or randomly placed SNPs, two stage or one stage studies (Hirschhorn and Daly, 2005). With most studies using a mixture of these methods, it remains to be seen if one method will stand out as the most useful. Appropriate correction of multiple testing is vital if genome-wide association studies are to be interpreted correctly. Furthermore, the analysis of genome-wide association data needs to account for population stratification because this is a confounding factor of potentially large effect (Freedman *et al.*, 2004; Marchini *et al.*, 2004). Further work is needed to iron

out these design and analysis issues before genome-wide association analysis can fully take flight.

The efficacy of whole genome association studies for the identification of polymorphisms underlying complex traits is a controversial issue (Terwilliger and Hiekkalinna, 2006). Results from candidate gene association studies often remain unreplicated, and there seems no reason to suggest that the situation will be different with genome-wide association studies.

Association mapping is based on the linkage disequilibrium between a marker locus and a quantitative trait nucleotide (QTN). The amount of linkage disequilibrium between two loci can be quantified using the r^2 formula. This is a function of the allele frequency at the two loci (for example a marker SNP and a QTN) and the degree of disequilibrium between the alleles. This measure of linkage disequilibrium is scaled by the allele frequency of the SNP within the population. The power to detect an association between a given SNP and a QTN is given by

$$\text{Power} \propto (r^2_{M,Q})nq^2 \tag{6.2.1}$$

where $r^2_{M,Q}$ is the extent of linkage disequilibrium between a marker (M) and a QTN (Q), n is the sample size and q^2 is the proportion of variance explained by the QTN (Wray, 2005). Therefore, the greater the correlation between the marker and the QTN (r^2) the greater the power to detect association. For a given amount of LD, r^2 can only be maximised when the frequency of the QTN is equal to that of the marker SNP. Thus, because the current set of marker SNPs have a minor allele frequency of approximately 5% or more, the success of genome-wide association studies hangs on the theory that loci underlying complex trait variation will be common within the population, the common-disease common

variant hypothesis (Blangero, 2004; Wang *et al.*, 2005; Wray, 2005). There has been some theoretical and empirical support (Reich and Lander, 2001) for this theory although others have suggested that a common-disease rare variant hypothesis is more likely (Pritchard, 2001; Pritchard and Cox, 2002). If the common-disease rare variant hypothesis is correct then several millions of SNP markers will be need to be typed across a huge number of individuals to identify the causal QTNs. For example, Table 6.2.1 shows that 6,852 individuals are needed to map a QTN that explains 0.01 of the phenotypic variance, assuming a type I error rate of 5×10^{-7} (the significance level for a Bonferroni correction following the testing of 1 million SNP markers) and a case-control study design. However, this number of individuals was calculated assuming that the marker SNP and the QTN were equally frequent within the population. If the minor allele frequency of the marker SNP is 0.05 and that of the QTN is 0.01, the maximum r^2 is 0.192 and the sample size required to identify the QTN with 80% power increases to 35,840. Whilst the allelic architecture will vary from trait to trait, it seems probable that the majority of traits will lie somewhere between the common-disease common variant and common-disease rare variant hypotheses and therefore, genomewide association will identify some, but not all, loci underlying a given complex trait.

There are many potential uses of genome-wide association data beyond simple association analysis (Gibbs and Singleton, 2006). These other applications include homozygosity mapping, direct detection of copy number variants (structural variation), genomic characterization of experimental species and different human populations, and admixture mapping.

Due to the flexibility of the data, the genome-wide association study is regarded to be one of the most exciting recent advances in the field. However, with many

investigators looking at ways to unravel the genetic nature of complex human phenotypes, new methods and technologies are appearing almost daily. One method that has caught the attention of many working within the field is the application and extension of gene expression studies for the analysis of complex phenotypes (King and Wilson, 1975).

6.2.3 Gene expression studies

The advent of microarray technology allowed large-scale studies of gene expression to take place (Rockman and Kruglyak, 2006). It can be argued that this is a more intuitive approach to identifying loci underlying complex trait variation because phenotypes are directly influenced by the products of genes (proteins), and only indirectly by the DNA sequence.

Jansen and Nap (2001) suggested an integration of the two approaches, where the levels of gene expression for multiple genes are treated as a quantitative trait in a traditional linkage analysis. Brem *et al.* (2002) were the first to apply this method, and they identified 570 expression QTL (eQTL) in *Saccharomyces cerevisiae* (yeast). Subsequently, many eQTL mapping studies have been carried out for a wide range of species, with many focussing on the mapping of eQTL in humans using either a genome-wide linkage (Schadt *et al.*, 2003; Monks *et al.*, 2004; Morley *et al.*, 2004), or genome-wide association approach (Cheung *et al.*, 2005; Stranger *et al.*, 2005). With traditional linkage analyses it is the volume of genetic marker data which makes the analysis difficult. With the genetic mapping of microarray data it is not only the number of genetic markers but also the sheer number of expression traits (typically thousands) which increase the computational intensity of the analysis. To maintain the correct significance level, a multiple testing correction must be applied that accounts for both the number of genetic markers and the multiple expression traits

(Kendziorski and Wang, 2006).

The rationale behind eQTL mapping studies is that the chromosomal region where sequence variation occurs accounts for some of the variation in the gene expression. Wayne and McIntyre (2002) carried out a QTL mapping experiment using *Drosophila* and subsequently used gene expression microarray technology to identify candidate genes within the QTL. This method greatly reduces the number of candidate genes within a given QTL and requires no knowledge on behalf of the investigator with regard to suitable candidate genes.

The genome-wide mapping of eQTL has the potential to identify gene-gene interactions, biochemical pathways and enables genetically alike individuals to be grouped together (Darvasi, 2003). The ability to classify individuals according to gene expression profiles for a given trait is potentially a significant step forward in our ability to identify genetically homogeneous subgroups of individuals or families for genetic linkage and/or association studies. In addition, gene expression profiles can be used to identify people who would benefit from a certain treatment and this would be an important step toward personalized medicines.

A problem associated with the analysis of microarray data is how best to handle those cases where the probe-set has reached the maximum measurement threshold. This may occur because a particular gene is been expressed to a high level, but can also arise due to operator error. Currently, the saturated probe-sets are removed from the analysis (Wang *et al.*, 2001) but this potentially introduces a bias and is likely to decrease statistical power. If these saturated chips are analysed as censored observations within a survival analysis framework, this bias and loss of power can be reduced. However, the analysis of genome-wide gene

expression data is already a computationally intensive task and using current survival analysis methods would significantly increase this. Survival analysis methods, similar to the linkage method suggested in Chapter 3, need to be developed which are computationally simple, thus enabling them to be applied to genome-wide gene expression data. These computationally simple methods would also be of benefit when analysing survival phenotypes using traditional genome-wide association or linkage data.

6.3 Conclusions

The use of age at onset data to identify loci underlying complex traits is an ever expanding field. The survival analysis methods which are used to analyse genetic data need to be modified to cope with the sheer volumes of data being generated in modern genetic studies. The incorporation of computationally efficient survival analysis methods into widely used genome analysis packages, such as MERLIN or SOLAR (Abecasis *et al.*, 2002; Duggirala *et al.*, 1997; Almasy and Blangero, 1998), would lead to an increase in the use of age at onset data. In an attempt to obtain a homogeneous sample of individuals for genetic analysis, more and more investigators will be using the age at onset of disease as a defining characteristic. Age at onset data has already proved the key to the identification of loci underlying several traits, both Mendelian and complex, and this seems set to continue.

Despite a few false dawns, the mapping of loci underlying complex traits is now a distinct possibility. Over the next five years we should see an increasing number of complex trait loci identified. By using linkage analysis to analyse families segregating narrowly-defined subsets of a disease, rare loci of large effect will be identified for many complex traits. The huge amount of phenotypic

information held on hundreds of thousands of individuals in population biobanks will enable large-scale genetic studies of complex traits and this should lead to the identification of many loci of medium effect size. The continued success of genome-wide association analysis will identify many of the common polymorphisms of medium to large effect which underpin complex trait variation. The recent advent of genome-wide gene expression data allows a more direct link between genotype and phenotype. The identification of eQTL that have different levels of expression in affected and unaffected individuals presents a new and exciting opportunity to identify loci underlying variation in complex traits.

There are still many obstacles to overcome. Methods need to be developed for including gene-gene and gene-environment interactions into gene mapping studies. The trend in genetics is to discount the environmental component of variation as if it were a stand alone component of variation outside of our interest (Hemminki *et al.*, 2006). However, as geneticists we should be paying great attention to environmental variation because of its interaction with genetic factors. If gene-gene and gene-environment interactions are to be successfully identified then we will need very large sample sizes, perhaps even bigger than those available in most biobanks. The release of genetic data into the public domain could create massive samples of phenotypically similar individuals. The Genetic Association Information Network (http://www.fnih.org/GAIN/) and The Wellcome Trust Case Control Consortium (http://www.wtccc.org.uk) have already committed themselves to the conditional public release of data. However, there are many ethical, political and legal issues to overcome before this becomes widespread.

Many geneticists are focussing on the problem of identifying the QTN of small effect which underlie variation in complex traits. These QTL are extremely

difficult to map using either a linkage or association approach, and the sample sizes required to map such loci are almost incomprehensible. However, whilst the identification of these small effect QTN may aid our understanding of the biology underlying the trait, these QTN will have little impact on an individual's risk of disease. Therefore, in terms of personalized medicine and the identification of people at high risk of common diseases, these loci are of little use. The only loci of real predictive value individually are those that confer a reasonable risk of disease, and it seems probable that these will be identified in the coming years. Therefore, the promised land of personalized medicine is perhaps not as far away as some would have us believe.

Bibliography

Abecasis, G., Cherny, S., Cookson, W. and Cardon, L. (2002). Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**: 97–101.

Abecasis, G., Cox, N., Daly, M., Kruglyak, L., Laird, N., Markianos, K. and Patterson, N. (2004). No bias in linkage analysis. *American Journal of Human Genetics* **75**: 722–723.

Alcais, A., Plancoulaine, S. and Abel, L. (2001). An autosomal-wide search for loci underlying wheezing age of onset in German asthmatic children identifies a new region of interest on 6q24-q25. *Genetic Epidemiology* **21 (Suppl 1)**: S168–S173.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**: 1198–1211.

Altmuller, J., Palmer, L., Fischer, G., Scherb, H. and Wjst, M. (2001). Genomewide scans of complex human diseases: True linkage is hard to find. *American Journal of Human Genetics* **69**: 936–950.

Altshuler, D., Hirschhorn, J., Klannemark, M., Lindgren, C., Vohl, M., Nemesch, J., Lane, C., Schaffner, S., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T., Daly, M., Groop, L. and Lander, E. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics* **26**: 76–80.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 1400 K Street, N.W., Washington, DC 20005, 4th edition.

Andersson, L. and Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics* **5**: 202–212.

Badano, J. and Katsanis, N. (2002). Beyond Mendel: An evolving view of human genetic disease transmission. *Nature Reviews Genetics* **3**: 779–789.

Balciuniene, J., Yuan, Q., Engström, C., Lindblad, K., Nylander, P., Sundvall, M., Schalling, M., Petterson, U., Adolfsson, R. and Jazin, E. (1998). Linkage analysis of candidate loci in families with recurrent major depression. *Molecular Psychiatry* **3**: 162–168.

Baret, P., Knott, S. and Visscher, P. (1998). On the use of linear regression and maximum likelihood for QTL mapping in half-sib designs. *Genetical Research* **72**: 149–158.

Bell, G., Horita, S. and Karam, J. (1984). A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33**: 176–183.

Bland, R., Newman, S. and Orn, H. (1986). Recurrent and nonrecurrent depression. *Archives of General Psychiatry* **43**: 1085–1089.

Blangero, J. (2004). Localization and identification of human quantitative trait loci: King Harvest has surely come. *Current Opinion in Genetics and Development* **14**: 233–240.

Botstein, D., White, R., Skolnick, M. and Davis, R. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**: 314–331.

Brem, R., Yvert, G., Clinton, R. and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.

Camp, N. and Cannon-Albright, L. (2005). Dissecting the genetic etiology of major depressive disorder using linkage analysis. *TRENDS in Molecular Medicine* **11**: 138–144.

Camp, N., Lowry, M., Richards, R., Plenk, A., Carter, C., Hensel, C., Abkevich, A., Skolnick, M., Shattuck, D., Rowe, K., Hughes, D. and Cannon-Albright, L. (2005). Genome-wide linkage analysis of extended Utah pedigrees identifies loci that influence recurrent, early-onset major depression and anxiety disorder. *American Journal of Medical Genetics* **135B**: 85–93.

Caprio, M., Fabbrini, E., Isidori, A., Aversa, A. and Fabbri, A. (2001). Leptin in reproduction. *TRENDS in Endocrinology & Metabolism* **12**: 65–72.

Cardon, L. and Bell, J. (2001). Association study designs for complex diseases. *Nature Reviews Genetics* **2**: 91–99.

Cheung, V., Spielman, R., Ewens, K., Weber, T., Morley, M. and Burdick, J. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**: 1365–1369.

Cichon, S., Schumacher, J., Muller, D., Hurter, M., Windemuth, C., Strauch, K., Hemmer, S., Schulze, T., Schmidt-Wolf, G., Albus, M., Borrmann-Hassenbach, M., Franzek, E., Lanczik, M., Fritze, J., Kreiner, R., Reuner, U., Weigelt, B., Minges, J., Lichtermann, D., Lerer, B., Kanyas, K., Baur, M., Wienker, T., Maier, W., Rietschel, M., Propping, P. and Nothen, M. (2001). A genome screen for genes predisposing to bipolar affective disorder detects a new susceptibility locus on 8q. *Human Molecular Genetics* **10**: 2933–2944.

Clerget-Darpoux, F., Bonaiti-Pellie, C. and Hochez, J. (1986). Effects of misspecifing genetic parameter in lod score analysis. *Biometrics* **42**: 393–399.

Collins, F. (1992). Positional cloning: Let's not call it reverse anymore. *Nature Genetics* **1**: 3–6.

Collins, F. (1995). Positional cloning moves from perditional to traditional. *Nature Genetics* **9**: 347–350.

Cooper, G., Ephross, S., Weinberg, C., Baird, D., Whelan, E. and Sander, D. (1999). Menstrual and reproductive risk factors for ischemic heart disease. *Epidemiology* **10**: 255–9.

Cordell, H. (2004). Bias toward the null hypothesis in model free linkage analysis is highly dependent on the test statistic used. *American Journal of Human Genetics* **74**: 1294–1302.

Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J. and Pericak-Vance, M. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**: 921–923.

Cornes, B., Medland, S., Ferreira, M., Morley, K., Duffy, D., Heijmans, B., Montgomery, G. and Martin, N. (2005). Sex-limited genome-wide linkage scan for body mass index in an unselected sample of 933 Australian twin families. *Twin Research and Human Genetics* **8**: 616–632.

Cox, D. (1972). Regression models in life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**: 187–220.

Cox, N., Reich, T., Rice, J., Elston, R., Schober, J. and Keats, B. (1989). Segregation and linkage analyses of bipolar and major depressive illnesses in multigenerational pedigrees. *Journal of Psychiatric Research* **23**: 109–123.

Damon, A. and Bajema, C. (1974). Age at menarche of recall after thirty-nine years. *Human Biology* **46**: 381–384.

Damon, A., Damon, S., Reed, R. and Valadian, I. (1969). Age at menarche of mothers and daughters, with a note on accuracy of recall. *Human Biology* **41**: 161–182.

Darvasi, A. (2003). Gene expression meets genetics. *Nature* **422**: 269–270.

Daw, E., Heath, S. and Wijsman, E. (1999). Multipoint oligogenic analysis of age-at-onset data with applications to Alzheimers disease pedigrees. *American Journal of Human Genetics* **64**: 839–851.

del P. Schneider, M., Strandberg, E., Ducrocq, V. and Roth, A. (2005). Survival analysis applied to genetic evaluation for female fertility in dairy cattle. *Journal of Dairy Science* **88**: 2253–2259.

Demerath, E., Towne, B., Chumlea, W., Sun, S., Czerwinski, S., Remsberg, K. and Siervogel, R. (2004). Recent decline in age at menarche: the Fels Longitudinal Study. *American Journal of Human Biology* **16**: 453–457.

deMuinich Keizer-Schrama, S. and Mul, D. (2001). Trends in pubertal development in Europe. *Human Reproduction Update* **7**: 287–291.

Diao, G. and Lin, D. (2005). Semiparametric methods for mapping quantitative trait loci with censored data. *Biometrics* **61**: 789–798.

Diao, G. and Lin, D. (2006). Semiparametric variance-component model for linkage and association analyses of censored trait data. *Genetic Epidemiology* **30**: 570–581.

Diao, G., Lin, D. and Zou, F. (2004). Mapping quantitative trait loci with censored observations. *Genetics* **168**: 1689–1698.

Duerr, R., Taylor, K., Brant, S., Rioux, J., Silverberg, M., Daly, M., Steinhart, A., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L., Kistner, E., Schumm, L., Lee, A., Gregersen, P., Barmada, M., Rotter, J., Nicolae, D. and Cho, J. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**: 1461–1463.

Duffy, D. (2002). Sib-pair version 0.99.9 [computer program]. Brisbane, Australia: Queensland Institute of Medical Research.

Duffy, D. (2006). An integrated genetic map for linkage analysis. *Behavior Genetics* **36**: 4–6.

Duggirala, R., Williams, J., Williams-Blangero, S. and Blangero, J. (1997). A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genetic Epidemiology* **14**: 987–992.

Dunnigan, M., Henderson, J., Hole, D. and Pelosi, A. (2004). Unexplained swelling symptoms in women (idiopathic oedema) comprise one component of a common polysymptomatic syndrome. *Quarterly Journal of Medicine* **97**: 755–764.

Dunnigan, M. and Pelosi, A. (1993). Familial idiopathic oedema in prepubertal children: a new syndrome. *Quarterly Journal of Medicine* **86**: 301–313.

Easton, D. (1999). How many more breast cancer predisposition genes are there? *Breast Cancer Research* **1**: 14–17.

Edenberg, H., Dick, D., Xuei, X., Tian, H., Almasy, L., Bauer, L., Crowe, R., Goate, A., Hesselbrock, V., Jones, K., Kwon, J., Li, T.-K., Nurnberger Jr, J., O'Connor, S., Reich, T., Rice, J., Schuckit, M., Porjesz, B., Foroud, T. and Begleiter, H. (2004). Variations in GABRA2, encoding the alpha2 subunit of the GABAa receptor, are associated with alcohol dependence and with brain oscillations. *American Journal of Human Genetics* **74**: 705–714.

Edwards, O. and Baylis, R. (1976). Idiopathic oedema of women. A clinical and investigative study. *Quarterly Journal of Medicine* **177**: 125–144.

Epstein, M., Lin, X. and Boehnke, M. (2003). A tobit variance-component method for linkage analysis of censored trait data. *American Journal of Human Genetics* **72**: 611–620.

Evans, D. and Medland, S. (2003). A note on including phenotypic information from monozygotic twins in variance components QTL linkage analysis. *Annals of Human Genetics* **67**: 613–617.

Falconer, D. and Mackay, T. (1996). *Introduction to quantitative genetics*. Pearson Education Limited, Edinburgh Gate, Harlow, Essex, CM20 2JE, England, 4th edition.

Frary, A., Nesbitt, T., Grandillo, S., Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K. and Tanksley, S. (2000). fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**: 85–88.

Freedman, M., Reich, D., Penney, K., McDonald, G., Mignault, A., Patterson, N., Gabriel, S., Topol, E., Smoller, J., Pato, C., Pato, M., Petryshen, T., Kolonel, L., Lander, E., Sklar, P., Henderson, B., Hirschhorn, J. and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics* **36**: 388–393.

Fridman, E., Pleban, T. and Zamir, D. (2000). A recombination hotspot delimts a wild-species quantitative trait locus for tomato sugar content to 484bp within an invertase gene. *Proceedings of the National Academy of Sciences* **97**: 4718–4723.

George, A., Visscher, P. and Haley, C. (2000). Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics* **156**: 2081–2092.

Gibbs, J. and Singleton, A. (2006). Application of genome-wide single nucleotide polymorphism typing: Simple association and beyond. *PLoS Genetics* **2**: e150.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J. and Thompson, R. (2002). ASReml user guide release 1.0.

Glazier, A., Nadeau, J. and Aitman, T. (2002). Finding genes that underlie complex traits. *Science* **298**: 2345–2349.

Goldgar, D. (2001). Major strengths and weaknesses of model-free methods. *Advances in Genetics* **42**: 241–251.

Gottsch, M., Clifton, D. and Steiner, R. (2006). Kisspeptin-GPR54 signalling in the neuroendocrine reproductive axis. *Molecular and Cellular Endocrinology* **254**: 97–102.

Greenberg, D., Abreu, P. and Hodge, S. (1998). The power to detect linkage in complex disease by means of simple lod-score analysis. *American Journal of Human Genetics* **63**: 870–879.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M. and Snell, R. (2002). Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**: 222–231.

Guo, Y., Shen, H., Xiao, P., Xiong, D., Yang, T., Guo, Y., Long, J., Recker, R. and Deng, H. (2006a). Genomewide linkage scan for quantitative trait loci underlying variation in age at menarche. *Journal of Clinical Endocrinology and Metabolism* **91**: 1009–1014.

Guo, Y., Xiong, D., Yang, T., Guo, Y., Recker, R. and Deng, H. (2006b). Polymorphisms of estrogen-biosynthesis genes CYP17 and CYP19 may influence age at menarche: a genetic association study in caucasian females. *Human Molecular Genetics* **15**: 2401–2408.

Haines, J., Hauser, M., Schmidt, S., Scott, W., Olson, L., Gallins, P., Spencer, K., Kwan, S., Noureddine, M., Gilbert, J., Scnetz-Boutaud, N., Agrawal, A., Postel, E. and Pericak-Vance, M. (2005). Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**: 419–421.

Haley, C. and Knott, S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.

Hall, J., Lee, M., Newman, B., Morrow, J., Anderson, L., Huey, B. and King, M.-C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**: 1684–1689.

Heath, A., Howells, W., Kirk, K., Madden, P., Bucholz, K., Nelson, E., Slutske, W., Statham, D. and Martin, N. (2001). Predictors of non-response to a questionnaire survey of a volunteer twin panel: findings from the Australian 1989 twin cohort. *Twin Research* **4**: 73–80.

Heath, S. C. (1997). Markov Chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61**: 748–760.

Hemminki, K., Bermejo, J. and Forsti, A. (2006). The balance between heritable and environmental aetiology of human disease. *Nature Reviews Genetics* **7**: 958–965.

Hirschhorn, J. and Daly, M. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**: 95–108.

Holmans, P., Zubenko, G., Crowe, R., DePaulo Junior, R., Scheftner, W., Weissman, M., Zubenko, W., Boutelle, S., Murphy-Eberenz, K., MacKinnon, D., McInnes, M., Marta, D., Adams, P., Knowles, J., Gladis, M., Thomas, J., Chellis, J., Miller, E. and Levinson, D. (2004). Genomewide significance linkage to recurrent, early-onset major depressive disorder on chromosome 15q. *American Journal of Human Genetics* **74**: 1154–1168.

Horikawa, Y., Oda, N., Cox, N., Li, X., Orho-Melander, M., Hara, M., Lindner, T., Mashima, H., Scwarz, P., del Bosque-Plata, L., Horikawa, Y., Oda, Y., Yoshiuchi, I., Colilla, S., Polonsky, K., Wei, S., Concannon, P., Iwasaki, N., Schulze, J., Baier, L., Bogardus, C., Groop, L., Boerwinkle, E., Hanis, C. and Bell, G. (2000). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics* **26**: 163–175.

Hoth, C., Milunsky, A., Lipsky, N., Sheffer, R., Clarren, S. and Baldwin, C. (1993). Mutations in the paired domain of the human PAX3 gene cause Klein-Waardenburg syndrome (WS-III) as well as Waardenburg syndrome type i (WS-I). *American Journal of Human Genetics* **52**: 455–462.

Hughes, I. and Kumanan, M. (2006). A wider perspective on puberty. *Molecular and Cellular Endocrinology* **254-255**: 1–7.

Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.-P., Belaiche, J.-P., Almerk, S., Tysk, C., O'Morain, C., Gassull, M., Binder, M., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macrykk, J., Colombel, J., Sahbatou, J. and Thomas, G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**: 599–603.

Huhtaniemi, I. (2002). LH and FSH receptor mutations and their effects on puberty. *Hormone Research* **57(suppl 2)**: 35–38.

Jansen, R. and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. *TRENDS in Genetics* **17**: 388–391.

Jeon, J., Carlborg, O., Tornsten, A., Giuffra, E., Amarger, V., Chardon, P., Andersson-Eklund, L., Andersson, K., Hansson, I., Lundstrom, K. and Andersson, L. (1999). A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nature Genetics* **21**: 157–158.

Kajiwara, K., Berson, E. and Dryja, T. (1994). Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science* **264**: 1604–1608.

Kaltiala-Heino, R., Kosunen, E. and Rimpela, M. (2003). Pubertal timing, sexual behaviour and self-reported depression in middle adolescence. *Journal of Adolescence* **26**: 531–545.

Kaltiala-Heino, R., Rimpella, M., Rissanen, A. and Rantanen, P. (2001). Early puberty and early sexual activity are associated with bulimic-type eating pathology in middle adolescence. *Journal of Adolescent Health* **28**: 346–352.

Kao, C.-H. (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics* **156**: 855–865.

Kaprio, J., Rimpela, A., Winter, T., Viken, R., Rimpela, M. and Rose, R. (1995). Common genetic influences on BMI and menarche. *Human Biology* **67**: 739–753.

Kendler, K., Gatz, M., Gardner, C. and Pederson, N. (2005). Age at onset and familial risk for major depression in a Swedish national sample. *Psychological Medicine* **25**: 217–232.

Kendziorski, C. and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome.* **17**: 509517.

Kessler, R., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K., Rush, J., Walters, E. and Wang, P. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of American Medical Association* **289**: 3095–3105.

Kessler, R., Berglund, P., Demler, O., Jin, R., Koretz, D. and Walters, E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replacation (NCS-R). *Archives of General Psychiatry* **62**: 593–602.

Kessler, R., McGonagle, K., Zhao, S., Nelson, C., Hughes, M., Eshleman, S., Wittchen, H.-U. and Kendler, K. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. *Archives of General Psychiatry* **51**: 8–19.

King, M. and Wilson, A. (1975). Evolution at two levels in humans and chimpanzees. *Science* **188**: 107116.

Kirk, K., Blomberg, S., Duffy, D., Heath, A., Owens, I. and Martin, N. (2001). Natural selection and quantitative genetics of life-history traits in western women: a twin study. *Evolution* **55**: 423–435.

Klein, J. and Moeschberger, M. (1999). *Survival Analysis: Techniques for censored and truncated data*. Statistics for Biology and Health, Springer-Verlag, New York, Inc, 175 Fifth Avenue, New York, NY, 10010, USA, 3 edition.

Klein, R., Zeiss, C., Chew, E., Tsai, J.-Y., Sackler, R., Haynes, C., Henning, A., SanGiovanni, J., Mane, S., Mayne, S., Bracken, M., Ferris, F., Ott, J., Barnstable, C. and Hoh, J. (2005). Complement factor H and polymorphism in age-related macular degeneration. *Science* **308**: 385–389.

Knott, S. (2005). Regression-based quantitative trait loci mapping: robust, efficient and effective. *Philosophical Transactions of the Royal Society B* **360**: 1435–1442.

Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S., Frigge, M., Thorgeirsson, T., Gulcher, J. and Steffansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics* **31**: 241–247.

Koo, M. and Rohan, T. (1997). Accuracy of short-term recall of age at menarche. *Annals of Human Biology* **24**: 61–64.

Kreidberg, J., Sariola, H., Loring, J., Maeda, M., Pelletier, J., Housman, D. and Jaenisch, R. (1993). WT-1 is required for early kidney development. *Cell* **74**: 679–691.

Kroenke, K. and Mangelsdorff, A. (1989). Common symptoms in ambulatory care: incidence, evaluation, therapy, and outcome. *American Journal of Medcine* **86**: 262–266.

Kruglyak, L., Daly, M., Reeve-Daly, M. and Lander, E. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics* **58**: 1347–1363.

Kupfer, D., Frank, E., Carpenter, L. and Neiswanger, K. (1989). Family history in recurrent depression. *Journal of Affective Disorders* **17**: 113–119.

Lander, E. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Lander, E. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Science (USA)* **84**: 2363–2367.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**: 241–247.

Lander, E. and Schork, N. (1994). Genetic dissection of complex traits. *Science* **265**: 2037–2048.

Lathrop, G., Lalouel, J., Julier, C. and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences* **81**: 3443–3446.

Layman, L., Lee, E., Peak, D., Namnoum, A., Vu, K., van Lingen, B., Gray, M., McDonough, P., Reindollar, R. and Jameson, J. (1997). Delayed puberty and hypogonadism caused by mutations in the follicle-stimulating hormone beta-subunit gene. *New England Journal of Medicine* **337**: 607–611.

Li, H. (2002). An additive genetic gamma frailty model for linkage analysis of disease with variable age at onset using nuclear families. *Lifetime data analysis* **8**: 315–334.

Loesch, D., Huggins, R., Rogucka, E., Hoang, N. and Hopper, J. (1995). Genetic correlates of menarcheal age: a multivariate twin study. *Annals of Human Biology* **22**: 479–490.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland.

Madgwick, P. and Goddard, M. (1989). Genetic and phenotypic parameters of longevity in Australian dairy cattle. *Journal of Dairy Science* **72**: 264–2632.

Maher, B., Marazita, M., Zubenko, W., Spiker, D., Giles, D., Kaplan, B. and Zubenko, G. (2002). Genetic segregation analysis of recurrent, early-onset major depression: Evidence for single major locus transmission. *American Journal of Medical Genetics* **114**: 214–221.

Marazita, M., Neiswanger, K., Cooper, M., Zubenko, G., Giles, D., Frank, E., Kupfer, D. and Kaplan, B. (1997). Genetic segregation analysis of early-onset recurrent unipolar depression. *American Journal of Human Genetics* **61**: 1370–1378.

Marchini, J., Cardon, L., Phillips, M. and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36**: 512–517.

Matthews, C., Borgato, S., Beck-Peccoz, P., Adams, M., Tone, Y., Gambino, G., Casagrande, S., Tedeschini, G., Benedetti, A. and Chatterjee, V. (1993). Primary amenorrhoea and infertility due to a mutation in the beta-subunit of follicle-stimulating hormone. *Nature Genetics* **5**: 83–86.

Matthews, C. and Chatterjee, V. (1997). Isolated deficiency of follicle-stimulating hormone re-visited. *New England Journal of Medicine* **337**: 642.

McGuffin, P., Knight, J., Breen, G., Brewster, S., Boyd, P. and Craddock, N. (2005). Whole genome linkage scan of recurrent depressive disorder from the Depression Network Study. *Human Molecular Genetics* **14**: 3337–3345.

Mendlewicz, J. and Baron, M. (1981). Morbidity risks in subtypes of unipolar depressive illness: differences between early and late onset forms. *British Journal of Psychiatry* **139**: 463–466.

Messager, S., Chatzidaki, E., Ma, D., Hendrick, A., Zahn, D., Dixon, J., Thresher, R., Malinge, I., Lomet, D., Carlton, M., Colledge, W., Caraty, A. and Aparicio, S. (2005). Kisspeptin directly stimulates gonadotropin-releasing hormone release via G protein-coupled receptor 54. *Proceedings of the National Academy of Sciences* **102**: 1761–1766.

Meuwissen, T., Veerkamp, R., Engel, B. and Brotherstone, S. (2002). Single and multitrait estimates of breeding values for survival using sire and animal models. *Animal Science* **75**: 15–24.

Meyer, J., Eaves, L., Heath, A. and Martin, N. (1991). Estimating genetic influences on the age-at-menarche: a survival analysis approach. *American Journal of Medical Genetics* **39**: 148–154.

Miki, Y., Swensen, J., Shattuckeidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Qingyun, L., Cochran, C., Bennett, M., Ding, W., Bell, R., Rosenthal, J., Hussey, C., Tran, T., McClure, M., Frye, C., Hattier, T., Phelps, R., Haugen-Strano, A., Katcher, H., Yakumo, K., Gholami, Z., Shaffer, D., Stone, S., Bayer, S., Wray, C., Bogdan, R., Dayananth, P., Ward, J., Tonin, P., Narod, S., Bristow, P., Norris, F., Helering, L., Morrison, P., Rosteck, P., Lai, M., Barrett, J., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A. and Skolnick, M. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**: 66–71.

Miller, S., Dykes, D. and Polesky, H. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research* **16**: 1215.

Missmer, S., Hankinson, S., Spiegelman, D., Barbieri, R., Malspeis, S., Willet, W. and Hunter, D. (2004). Reproductive history and endometriosis among premenopausal women. *Obstetrics and Gynecology* **104**: 965–974.

Monaco, A., Neve, R., Colletti-Feener, C., Bertelson, C., Kurnit, D. and Kunkel, L. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323**: 646–650.

Monks, S., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J., Sachs, A. and Schadt, E. (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics* **75**: 1094–1105.

Morabia, A., Costanza, M. and The World Health Organization Collaborative Study of Neoplasia and Steroid Contraceptives (1998). International variability in ages at menarche, first livebirth and menopause. *American Journal of Epidemiology* **148**: 1195–1205.

Moreno, C., Elsen, J., Le Roy, P. and Ducrocq, V. (2005). Interval mapping methods for detecting QTL affecting survival and time-to-event phenotypes. *Genetical Research* **85**: 139–149.

Morgante, M. and Salamini, F. (2003). From plant genomics to breeding practice. *Current Opinion in Biotechnology* **14**: 214–219.

Morley, M., Moloney, C., Weber, T., Devlin, J., Ewens, K., Spielman, R. and Cheung, V. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743–747.

Mukhopadhyay, I., Feingold, E. and Weeks, D. (2004). No 'bias' toward the null hypothesis in most conventional multipoint nonparametric linkage analysis. *American Journal of Human Genetics* **75**: 716–718.

Murray, C. and Lopez, A. (1996). Evidence-based health policy - lessons from the Global Burden of Disease Study. *Science* **274**: 740–743.

Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321–324.

Naves, M., Diaz-Lopez, J., Gomez, C., Rodriguez-Rebollar, A. and Cannata-Andia, J. (2005). Determinants of incidence of osteoporotic fractures in the female Spanish population older than 50. *Osteoporosis International* **16**: 2013–2017.

Neale, M., Boker, S., Xie, G. and Maes, H. (2002). Mx: Statistical modeling. VCU Box 900126, Richmond, VA 23298: Department of Psychiatry. 6th Edition.

Nezer, C., Moreau, L., Brouwers, B., Coppieters, W., Detilleux, J., Hanset, R., Karim, L., Kvasz, A., Leroy, P. and Georges, M. (1999). An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature Genetics* **21**: 155–156.

Nurnberger, J., Jr, Foroud, T., Flury, L., Su, J., Meyer, E., Hu, K., Crowe, R., Edenberg, H., Goate, A., Bierut, L., Reich, T., Schuckit, M. and Reich, W. (2001). Evidence for a locus on chromosome 1 that influences vulnerability to alcoholism and affective disorder. *American Journal of Psychiatry* **158**: 718–724.

Nyholt, D., Morley, K., Ferreira, M., Medland, S., Boomsma, D., Heath, A., Merikangas, K., Montgomery, G. and Martin, N. (2005). Genomewide significant linkage to migrainous headache on chromosome 5q21. *American Journal of Human Genetics* **77**: 500–512.

O'Connell, J. and Weeks, D. (1998). Pedcheck: A program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics* **63**: 259–266.

Ogura, Y., Bonen, D., Inohara, N., Nicolae, D., Chen, F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerrk, R., Achkar, J.-P., Brant, S., Bayless, T., KirschnerI, B., Hanauer, S., Nunez, G. and Cho, J. (2001). A frame shift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**: 603–606.

Ong, K., Ahmed, M. and Dunger, D. (2006). Lessons from large population studies on timing and tempo of puberty (secular trends and relation to body size): The European trend. *Molecular and Cellular Endocrinology* **254-255**: 8–12.

Ott, J. (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics* **51**: 283–290.

Ott, J., Terwilliger, J. and Xie, X. (1992). Determining the informativeness of untyped individuals in a pedigree analysis. *American Journal of Human Genetics* **51**: A197.

Ozaki, K., Inoue, K., Sato, H., Iida, A., Ohnishi, Y., Sekine, A., Sato, H., Odashiro, K., Nobuyoshi, M., Hori, M., Nakamura, Y. and Tanaka, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32**: 650–654.

Pal, D., Durner, M. and Greenberg, D. (2001). Effect of misspecification of gene frequency on the two-point lod score. *European Journal of Human Genetics* **9**: 855–859.

Pankratz, V., de Andrade, M. and Therneau, T. (2005). Random-effects Cox proportional hazards model: Genetic variance components methods for time-to-event data. *Genetic Epidemiology* **28**: 97–109.

Pei, Y. (2003). Molecular genetics of autosomal dominant polycystic kidney disease. *Clinical and Investigative Medicine* **26**: 252–8.

Pelosi, A., Sykes, R., Lough, J., Muir, W. and Dunnigan, M. (1986). A psychiatric study of idiopathic oedema. *The Lancet* **2**: 999–1002.

Prentice, R. and Gloeckler, L. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**: 57–67.

Pritchard, J. (2001). Are rare variants responsible for susceptibility to complex disease? *American Journal of Human Genetics* **69**: 124–137.

Pritchard, J. and Cox, N. (2002). The allelic architecture of human disease genes: common-disease-common variant...or not? *Human Molecular Genetics* **11**: 2417–2423.

Purcell, S., Cherny, S. and Sham, P. (2003). Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**: 149–150.

R Development Core Team (2005). R: A language and environment for statistical computing. Http://www.R-project.org.

Rebai, A. (1997). Comparison of methods of regression interval mapping in QTL analysis with non-normal traits. *Genetical Research* **69**: 69–74.

Rebai, A., Goffinet, B. and Mangin, B. (1995). Comparing power of different methods for QTL detection. *Genetics* **51**: 87–99.

Reich, D. and Lander, E. (2001). On the allelic spectrum of human disease. *TRENDS in Genetics* **17**: 502–509.

Rice, J., Saccone, N. and Rasmussen, E. (2001). Definition of the phenotype. *Advances in Genetics* **42**: 69–76.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**: 1016–1022.

Risch, N. (2000). Searching for genetic determinants in the new millenium. *Nature* **405**: 847–856.

Risch, N. and Giuffra, L. (1992). Model misspecification and multipoint linkage analysis. *Human Heredity* **42**: 77–92.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**: 1616.

Roche, A. (1992). *Growth, maturation and body composition: The Fels Longitudinal Study*. Cambridge: Cambridge University Press.

Rockman, M. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics* **7**: 862–872.

Rothenbuhler, A., Fradin, D., Heath, S., Lefevre, H., Bouvattier, C., Lathrop, M. and Bougneres, P. (2006). Weight-adjusted genome scan analysis for mapping quantitative trait loci for menarchal age. *Journal of Clinical Endocrinology and Metabolism* **91**: 3534–3537.

Schadt, E., Monks, S., Drake, T., Lusis, A., Che, N., Colinayo, V., Ruff, T., Milligan, S., Lamb, J., Cavet, G., Linsley, P., Mao, M., Stoughton, R. and Friend, S. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.

Schork, N. and Greenwood, T. (2004). Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. *American Journal of Human Genetics* **74**: 306–316.

Scott, W., Grubber, J., Conneally, P., Small, G., Hulette, C., Rosenberg, C., Saunders, A., Roses, A., Haines, J. and Pericak-Vance, M. (2000). Fine mapping of the chromosome 12 late-onset Alzheimer disease locus: potential genetic and phenotypic heterogeneity. *American Journal of Human Genetics* **66**: 922–932.

Scriver, C. and Waters, P. (1999). Monogenic traits are not simple: lessons from phenylketonuria. *TRENDS in Genetics* **15**: 267–272.

Seaton, G., Haley, C., Knott, S., Kearsey, M. and Visscher, P. (2002). QTL Express: Mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics* **18**: 339–340.

Self, S. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**: 605–610.

Serretti, A., Macciardi, F., Cusin, C., Lattuada, E., Souery, D., Lipp, O., Mahieu, B., Van Broeckhoven, C., Blackwood, D., Muir, W., Aschauer, H., Heiden, A., Ackenheil, M., Fuchshuber, S., Raeymaekers, P., Verheyen, G., Kaneve, R., Jablensky, A., Papadimitriou, G., Dikeos, D., Stefanis, C., Smeraldi, E. and Mendlwicz, J. (2000). Linkage of mood disorders with D2, D3 and TH genes: A multicenter study. *Journal of Affective Disorders* **58**: 51–61.

Sham, P., Cherny, S., Purcell, S. and Hewitt, J. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics* **66**: 1616–1630.

Sharp, L., Cardy, A., Cotton, S. and Little, J. (2004). CYP17 gene polymorphisms: prevalence and associations with hormone levels and related factors. a HuGE review. *American Journal of Epidemiology* **160**: 729–740.

Sieberts, S., Broman, K. and Gudbjartsson, D. (2004). 'bias towards the null' means reduced power. *American Journal of Human Genetics* **75**: 720–722.

Smyth, D., Cooper, J., Bailey, R., Field, S., Burren, O., Smink, L., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D., Savage, D., Walker, N., Clayton, D. and Todd, J. (2006). A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature Genetics* **38**: 617–619.

Sobel, E., Sengul, H. and Weeks, D. (2001). Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Human Heredity* **52**: 121–131.

Stavrou, I., Zois, C., Ioannidis, J. and Tsatsoulis, A. (2002). Association of polymorphisms of the oestrogen receptor alpha gene with the age at menarche. *Human Reproduction* **17**: 1101–1105.

Strachan, T. and Read, A. (2003). *Human Molecular Genetics 3*. Garland Science Inc., New York.

Stram, D. and Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**: 1171–1177.

Stranger, B., Forrest, M., Clark, A., Minichiello, M., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S., Tavare, S., Deloukas, P. and Dermitzakis, E.

(2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics* **1**: e78.

Streeten, D. (1978). Idiopathic oedema: pathogenesis, clinical features and treatment. *Metabolism* **27**: 353–383.

Sullivan, P., Neale, M. and Kendler, K. (2000). Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry* **157**: 1552–1562.

Symons, R., Daly, M., Fridlyand, J., Speed, T., Cook, W., Gerondakis, S., Harris, A. and Foote, S. (2002). Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in Eμ-v-abl transgenic mice. *Proceedings of the National Academy of Sciences* **99**: 11299–11304.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research* **17**: 6463–6471.

Tena-Sempere, M. (2006). The roles of kisspeptins and G protein-coupled receptor-54 in pubertal development. *Current Opinion in Pediatrics* **18**: 442447.

Terwilliger, J. and Hiekkalinna, T. (2006). An utter refutation of the 'fundamental theorem of the hapmap'. *European Journal of Human Genetics* **14**: 426–437.

Terwilliger, J. and Ott, J. (1994). *Handbook of Human Genetic Linkage*. The John Hopkins University Press, 2715 North Charles Street, Baltimore, Maryland, 21218-4363.

Thapar, A. and McGuffin, P. (1994). A twin study of depressive symptoms in childhood. *British Journal of Psychiatry* **165**: 259–265.

The International HapMap Consortium (2003). The international hapmap project. *Nature* **426**: 789–796.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**: 1299–1320.

The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–912.

The International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.

Therneau, T. (2003). On mixed-effect Cox models, sparse matrices, and modeling data from large pedigrees. http://mayoresearch.mayo.edu/mayo/research/biostat/upload/kinship.pdf.

Thomas, D., Haile, R. and Duggan, D. (2005). Recent developments in genomewide association scans: A workshop summary and review. *American Journal of Human Genetics* **77**: 337–345.

Thorn, G. (1968). Approach to the patient with 'idiopathic edema' or 'periodic swelling'. *Journal of the American Medical Association* **206**: 333–338.

Towne, B., Czerwinski, S., Demerath, E., Blangero, J., Roche, A. and Siervogel, R. (2005). Heritability of age at menarche in girls from the Fels Longitudinal Study. *American Journal of Physical Anthropology* **128**: 210–219.

Treloar, S., Hadfield, R., Montgomery, G., Lambert, A., Wicks, J., Barlow, D., O'Connor, D., Kennedy, S. and the International Endogene Study Group. (2002). The International Endogene Study: a collection of families for genetic research in endometriosis. *Fertility and Sterility* **78**: 679–685.

Treloar, S. and Martin, N. (1990). Age at menarche as a fitness trait: nonadditive genetic variance detected in a large twin sample. *American Journal of Human Genetics* **47**: 137–148.

Treloar, S., Wicks, J., Nyholt, D., Montgomery, G., Bahlo, M., Smith, V., Dawson, G., Mackay, I., Weeks, D., Bennett, S., Carey, A., Ewen-White, K., Duffy, D., O'Connor, D., Barlow, D., Martin, N. and Kennedy, S. (2005). Genomewide linkage study in 1,176 affected sister pair families identifies a significant susceptibility locus for endometriosis on chromosome 10q26. *American Journal of Human Genetics* **77**: 365–376.

Tsui, L.-C., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, J., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., Zsiga, M., Markiewicz, D., Aktos, G., Brown, V., Helms, C., Gravius, T., Parker, C., Rediker, K. and Donis-Keller, H. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic dna marker. *Science* **230**: 1054–1057.

van den Berg, S., Setiawan, A., Bartels, M., Polderman, T., van der Vaart, A. and Boomsma, D. (2006). Individual differences in puberty onset in girls: Bayesian estimation of heritabilities and genetic correlations. *Behavior Genetics* **36**: 261–270.

van Eerdewegh, P., Little, R., Dupuis, J., Del Mastro, R., Falls, K., Simon, J., Torrey, D., Pandit, S., McKenny, J., Braunschweiger, K., Walsh, A., Liu, Z., Hayward, B., Folz, C., Manning, S., Bawa, A., Saracino, L., Thackston, M., Benchekroun, Y., Capparell, N., Wang, M., Adair, R., Feng, Y., Dubois, J., Fitzgerald, M., Huang, H., Gibson, R., Allen, K., Pedan, A., Danzig, M., Umland, S., Egan, R., Cuss, F., Rorke, S., Clough, J., Holloway, J., Holgate, S. and Keith, T. (2002). Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* **418**: 426–430.

Velie, E., Nechuta, S. and Osuch, J. (2006). Lifetime reproductive and anthropometric risk factors for breast cancer in postmenopausal women. *Breast Disease* **24**: 17–35.

Venken, T., Claes, S., Sluijs, S., Paterson, A., van Duijn, C., Adolfsson, R., Del-Favero, J. and Van Broeckhoven, C. (2005). Genomewide scan for affective disorder susceptibility loci in families of a northern Swedish isolated population. *American Journal of Human Genetics* **76**: 237–248.

Visscher, P., Haley, C., Heath, C., Muir, W. and Blackwood, D. (1999). Detecting QTLs for uni- and bipolar disorder using a variance component method. *Psychiatric Genetics* **9**: 75–84.

Visscher, P., Haley, C. and Knott, S. A. (1996). Mapping QTLs for binary traits in backcross and F2 populations. *Genetical Research* **68**: 55–63.

Visscher, P. and Wray, N. (2004). Conventional multipoint nonparametric linkage analysis is not necessarily inherently biased. *American Journal of Human Genetics* **75**: 718–720.

Wang, W., Barratt, B., Clayton, D. and Todd, J. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics* **6**: 109–118.

Wang, X., Ghosh, S. and Guo, S.-W. (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research* **29**: e75.

Wayne, M. and McIntyre, L. (2002). Combining mapping and arraying: An approach to candidate gene identification. *Proceedings of the National Academy of Sciences* **99**: 14903–6.

Weatherall, D. (2001). Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nature Reviews Genetics* **2**: 245–255.

Weber, J. and May, P. (1989). Abundant class of human polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44**: 388–396.

Wei, L., Lin, D. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association* **84**: 1065–1073.

Weissman, M., Gershon, E., Kidd, K., Prusoff, B., Leckman, J., Dibble, E., Hamovit, J., Thompson, W., Pauls, D. and Guroff, J. (1984). Psychiatric disorders in the relatives of probands with affective disorders. *Archives of General Psychiatry* **41**: 13–21.

Weissman, M., Wickramaratne, P., Merikangas, K., Leckman, J., Prusoff, B., Caruso, K., Kidd, K. and Gammon, G. (2005). Families at high and low risk for depression: a 3-generation study. *Archives of General Psychiatry* **62**: 29–36.

Wickramaratne, P., Warner, V. and Weissman, M. (2000). Selecting early onset MDD probands for genetic studies: results from a longitudinal high-risk study. *American Journal of Medical Genetics* **96**: 93–101.

Woo, S., Lisdkey, A., Guttler, F., Chandra, D. and Robson, K. (1983). Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. *Nature* **306**: 151–155.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., Barfoot, R., Hamoud, R., Patel, S., Rice, C., Biggs, P., Hashim, Y., Smith, A., Connor, F., Arason, A., Gudmundsson, J., Ficenec, D., Kelsell, D., Ford, D., Tonin, P., Bishop, D., Spurr, N., Ponder, B., Eeles, R., Peto, J., Devilee, P., Corneliese, C., Lynch, H., Narod, S., Lenoir, G., Egilsson, V., Barkadottir, R., Easton, D., Bentley, D., Futreal, P., Ashworth, A. and Stratton, M. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**: 789–792.

Wooster, R., Neuhausen, S., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D., Fields, P., Marshall, G., Narod, S., Lenoir, G., Lynch, H., Feunteun, J., Devilee, P., Cornelisse, C., Menko, F., Daly, P., Ormiston, W., McManus, R., Pye, C., Lewis, C., Cannon-Albright, L., Peto, J., Ponder, A., Skolnick, M., Easton, D., Goldgar, D. and Stratton, M. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**: 2088–2090.

Wray, N. (2005). Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Research and Human Genetics* **8**: 87–94.

Wright, M., de Geus, E., Ando, J., Luciano, M., Posthuma, D., Ono, Y., Hansell, N., van Baal, C., Hiraishi, K., Hasegawa, T., Smith, G., Geffen, G., Geffen, L., Kanba, S., Miyake, A., Martin, N. and Boomsma, D. (2001). Genetics of cognition: outline of a collaborative twin study. *Twin Research* **4**: 48–56.

Xita, N., Tsatsoulis, A., Stavrou, I. and Georgiou, I. (2005). Association of SHBG gene polymorphism with menarche. *Molecular Human Reproduction* **11**: 459–462.

Xue, X. and Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival times. *Lifetime Data Analysis* **2**: 277–289.

Yazdi, M., Visscher, P., Ducrocq, V. and Thompson, R. (2002). Heritability, reliability of genetic evaluations and response to selection in proportional hazards models. *Journal of Dairy Science* **85**: 1563–1577.

Yu, W., Kimura, M., Walczewska, A., Karanth, S. and McCann, S. (1997). Role of leptin in hypothalamic-pituitary function. *Proceedings of the National Academy of Sciences* **94**: 1023–1028.

Zhao, J. (2005). Mixed effects Cox models of alcohol dependence in extended families. *BMC Genetics* **6**: S127.

Zhu, G., Duffy, D., Eldridge, A., Grace, M., Mayne, C., OGorman, L., Aitken, J., Neale, M., Hayward, N., Green, A. and Martin, N. (1999). A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A: A maximum-likelihood combined linkage and association analysis in twins and their sibs. *American Journal of Human Genetics* **65**: 483–492.

Zhu, G., Evans, D., Duffy, D., Montgomery, G., Medland, S., Gillespie, N., Ewen, K., Jewell, M., Liew, Y., Hayward, N., Sturm, R., Trent, J. and Martin, N. (2004). A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q. *Twin Research* **7**: 197–210.

Zill, P., Engel, R., Baghai, T., Juckel, G., Frodl, T., Müller-Siechender, F., Zwanzger, P., Schüle, C., Minov, C., Behrens, S., Rupprecht, R., Hegerl, U., Möller, H. and Bondy, B. (2002). Identification of a naturally occurring polymorphism in the promoter region of the norepinephrine transporter and analysis in major depression. *Neuropsychopharmacology* **26**: 489–493.

Zubenko, G., Hughes III, H., Mahor, B., Stiffler, J., Zubenko, W. and Marazita, M. (2002). Genetic linkage of region containing the CREB1 gene to depressive disorders in women from families with recurrent, early-onset, major depression. *American Journal of Medical Genetics* **114**: 980–987.

Zubenko, G., Maher, B., Hughes III, H., Zubenko, W., Stiffler, J., Kaplan, B. and Marazita, M. (2003). Genome-wide linkage survey for genetic loci that influence the development of depressive disorders in families with recurrent, early onset, major depression. *American Journal of Medical Genetics* **123B**: 1–18.